Molecular Characterization of Mutant Germplasm

A Manual

Prepared by the
Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture

Plant Breeding and Genetics Laboratory, Seibersdorf, 2014

FOREWORD

Plant biotechnology applications must not only respond to the challenges of improving food security and fostering socio-economic development, but in doing so, promote the conservation, diversification and sustainable use of plant genetic resources for food and agriculture. Today the biotechnology toolbox available to plant breeders offers many new possibilities for accelerating the breeding process, and increasing productivity, crop diversification and production, while developing a more sustainable agriculture. The early versions of this manual provided a companion to training courses on plant mutant germplasm characterization. As such, the content was tailored to the curricula of the course. It has now developed to include new technologies as they emerge in providing a contemporary tool kit for genotypic analysis and selection in plant breeding and genetics.

The first print of this manual on selected molecular marker techniques was prepared using the hand-outs and other materials distributed to participants of the FAO/IAEA Interregional Training Course on "Mutant germplasm characterisation using molecular markers". The course was hosted by the Joint FAO/IAEA Programme of Nuclear Techniques in Food and Agriculture at the Plant Breeding and Genetics Laboratory (PBGL, formerly the Plant Breeding Unit) of the Agriculture and Biotechnology Laboratory at the IAEA Laboratories in Seibersdorf, Austria, in 2001. Messrs J. Bennetzen (USA), K. Devos (UK), G. Kahl (Germany), U. Lavi (Israel), M. Mohan (ICGEB) and S. Nielen (FAO/IAEA) contributed protocols to the first print version. These contributions and others were formally compiled into the first early editions of the manual by Messrs P. Gustafson (USA), B. Forster (UK, currently head of PBGL), M. Gale (UK), R. Adlam (UK), M. Maluszynski and S. Nielen of the Joint Programme. Their efforts in establishing this manual are deeply appreciated. In later editions, J Fernandez-Manjarres (Colombia) provided the section on population genetics, and Plant Breeding and Genetics Section Head Pierre Lagoda provided the protocol on multivariate analysis. While this series of courses ended in 2007, there has been a continual demand from trainees for a codified set of standard protocols, and so the Plant Breeding and Genetics Laboratory (PBGL) has continued adapting this book by incorporating new protocols with the aim of assisting Member States in the appropriate application of molecular tools with minimal costs. These include protocols for TILLING/Ecotilling, DNA quantification, low-cost and low toxicity DNA extraction, alternative enzymology for enzymatic mismatch cleavage (new in 2013), and methods for rapid bench-top purification of single-strand-specific nucleases used in mutation discovery assays (new in 2014). Of note in this 2014 edition is the successful implementation of low-cost and non-toxic DNA extraction methods developed by the PBGL and first delivered to the Member States in the 2013 edition of this manual. Because these methods have been successfully used at 4 training courses in 2013 and used in 20 different crops, the more traditional (and toxic) method of DNA extraction using organic phase separation has been deleted from this manual. Particular thanks for work on this recent edition go to PBGL staff Owen Huynh,, Joanna Jankowicz-Cieslak, and Bradley Till.

We strive to improve the manual with each edition. We very much appreciate feedback, suggestions and comments, which could further improve and enrich the contents of this

manual. Correspondence should be addressed directly to Mr. PJL Lagoda, Head of Plant Breeding and Genetics Section, Joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, P.O. Box 100, Vienna, Austria, Telephone: +43 1 2600 21626; email *P.Lagoda@iaea.org*.

A hard copy with attached CD-ROM will be distributed, free of charge, to interested scientists from FAO and IAEA Member States. Requests for the manual should be sent to Ms. K. Allaf, Plant Breeding and Genetics Section, Joint FAO/IAEA Division of Nuclear Application in Agriculture, P.O. Box 100, Vienna, Austria, Telephone: +43 1 2600 21621 or by email: *K.Allaf@iaea.org*.

LIST OF ACRONYMS

AFLP	**A**mplified **F**ragment **L**ength **P**olymorphism
CAPS	**C**leaved **A**mplified **P**olymorphic **S**equences
CJE	**C**elery **J**uice **E**xtract
EST	**E**xpressed **S**equence **T**ag
IPCR	**I**nverse **P**olymerase **C**hain **R**eaction
IRAP	**I**nter-**R**etrotransposon **A**mplified **P**olymorphism
ISSR	**I**nter-**S**imple **S**equence **R**epeat amplification
PCR	**P**olymerase **C**hain **R**eaction
RAPD	**R**andom **A**mplified **P**olymorphic **D**NA
REMAP	**Re**trotransposon-**M**icrosatellite **A**mplified **P**olymorphism
RFLP	**R**estriction **F**ragment **L**ength **P**olymorphism
SCAR	**S**equence **C**haracterized **A**mplified **R**egion
SNP	**S**ingle **N**ucleotide **P**olymorphism
SSCP	**S**ingle **S**tranded **C**onformation **P**olymorphism
SSR	**S**imple **S**equence **R**epeat
STS	**S**equence **T**agged **S**ite
TILLING	**T**argeting **I**nduced **L**ocal **L**esions **IN** **G**enomes
NGS	**N**ext **G**eneration **S**equencing

TABLE OF CONTENTS

FOREWORD I

LIST OF ACRONYMS III

TABLE OF CONTENTS IV

1. INTRODUCTION TO MOLECULAR MARKERS 1-1

1.1. Use of molecular markers: A cautionary tale 1-2

1.1.1. An example of how not to use molecular markers. 1-2

1.1.2. An example of efficient application of markers 1-3

1.2. A Summary of Marker Techniques 1-4

1.3. Ideal genetic markers 1-4

1.4. Marker application suitability 1-5

1.5. Implementation 1-8

1.6. Requirements 1-8

1.7. Comparison of different marker systems 1-9

2. LOW COST DNA EXTRACTION WITHOUT TOXIC ORGANIC PHASE SEPARATION 2-1

2.1. Materials 2-1

2.2. Solutions to Prepare 2-3

2.3. Methods (for centrifuge tubes) 2-3

2.4. Example Data 2-6

2.5. Conclusions 2-8

3. DNA QUANTIFICATION 3-1

3.1. Protocol for gel electrophoresis 3-1

3.1.1. Preparation of DNA concentration standards. 3-1

3.1.2. Preparing agarose gels. 3-2

3.1.3. Preparing samples for loading into gels. 3-2

3.1.4. Running the gel 3-2

3.1.5. Photographing the gel 3-3
3.2. Quantification of DNA using image analysis software 3-3

4. RESTRICTION ENZYME DIGEST 4-1

5. FINDING CANDIDATE GENES AND PRIMER DESIGN FOR MOLECULAR TESTING: AN EXAMPLE FROM THE ANNOTATED *SORGHUM BICOLOR* GENOME 5-1

5.1. Overview 5-1

6. RFLP 6-1

6.1. Protocol 6-2
6.1.1. Agarose gel electrophoresis 6-2
6.1.2. Southern blotting and hybridization 6-4
6.1.3. Labelling the probe and dot blot/quantification 6-7
6.2. Hybridisation 6-9
6.2.1. Washing method 6-9
6.2.2. Detection 6-10
6.3. Membrane rehybridisation method 6-12
6.4. References 6-12
6.5. Reagents needed 6-12

7. SSR 7-1

7.1. Protocol 7-2
7.1.1. PCR reaction mix 7-2
7.1.2. PCR amplification 7-3
7.1.3. Separation of the amplification products in agarose gel 7-3
7.1.4. Denaturing gel electrophoresis 7-4
7.1.5. Assembling the glass plate sandwich 7-4
7.1.6. Casting gel 7-5
7.2. Setting up the operation 7-5
7.3. Polyacrylamide gel running conditions 7-6
7.4. Silver-staining 7-6
7.5. References 7-7

7.6. Reagents needed 7-8

8. ISSR 8-1

8.1. Protocol 8-1

8.1.1. Prepare 20µl reaction mix 8-2

8.1.2. PCR amplification 8-2

8.1.3. Separation and visualization of the amplification products 8-2

8.1.4. Gel running conditions 8-3

8.1.5. Silver-staining 8-3

8.2. Primers available at Plant Breeding & Genetics Laboratory (FAO/IAEA) 8-3

8.3. References 8-4

8.4. Reagents needed 8-4

9. AFLP 9-1

9.1. Protocol 9-2

9.1.1. Restriction of genomic DNA and ligation of adapters to the DNA fragments 9-2

9.1.2. Pre-amplification 9-3

9.1.3. PCR pre-amplification 9-3

9.1.4. Check-step 9-3

9.1.5. Selective pre-amplification 9-4

9.1.6. PCR mix for selective amplification, products to be visualized on PAGE 9-5

9.1.7. PCR profile for Selective amplification, products to be visualised on PAGE. 9-5

9.1.8. Polyacrylamide Gel Electrophoresis (PAGE) 9-5

9.1.9. Silver staining of PAG 9-6

9.1.10. PCR mix for selective amplification, products to be visualized on an automated DNA analyser 9-6

9.1.11. PCR profile for selective amplification, products to be visualized on an automated DNA analyser 9-6

9.1.12. Electrophoresis using an automated DNA analyser 9-6

9.1.13. Production of single primer, linear PCR products 9-7

9.1.14. PCR amplification to produce single stranded DNA 9-7

9.2. Required enzymes and primer sequences for AFLP assays 9-8

9.2.1. Restriction enzymes 9-8

9.3. Preparation of adapters 9-8
9.4. Reagents needed 9-8
9.5. Sequence information of adapters and primers used for AFLP 9-9
9.6. References 9-10

10. REMAP & IRAP 10-1

10.1. Protocol 10-1
10.1.1. Prepare a 50µl reaction mix 10-2
10.1.2. PCR amplification 10-3
10.1.3. Separation and visualization of the amplification products 10-3
10.2. References 10-4
10.3. Reagents needed 10-4

11. SINGLE NUCLEOTIDE POLYMORPHISMS (SNPS) 11-1

11.1. References 11-2

12. TILLING 12-1

12.1. Protocol 12-1
12.1.1. PCR reaction with IRDye-labeled primers 12-1
12.1.2. Heteroduplex digestion, preparation of Sephadex spin plates 12-2
12.1.3. Agarose gel analysis of enzymatic mismatch cleavage, and sample purification 12-4
12.1.4. Sample purification and volume reduction 12-5
12.1.5. Preparing, loading, and running LI-COR gels 12-6
12.1.6. Data Analysis 12-8
12.2. Computation tools 12-8
12.2.1. Selecting the best region to screen and designing primers 12-8
12.3. Data analysis 12-10
12.4. Additional info 12-13
12.4.1. List of consumables and equipment 12-13
12.5. Frequently asked questions 12-15
12.6. Additional protocols 12-16
12.6.1. Sequencing 12-16

12.7. EMS mutagenesis of Arabidopsis seed 12-18
12.7.1. Materials 12-18
12.7.2. Standard size batch 12-18
12.7.3. A note on technique 12-19
12.7.4. DNA extraction 12-19
12.8. References 12-20

13. ALTERNATIVE ENZYMOLOGY FOR MISTMATCH CLEAVAGE FOR TILLING AND ECOTILLING: EXTRACTION OF ENZYMES FROM WEEDY PLANTS 13-1

13.1. Objective 13-1
13.2. Materials 13-1
13.3. Methods 13-2
13.3.1. Enzyme extraction 13-2
13.3.2. Concentration of enzyme extractions 13-3
13.3.3. Test of Mismatch Cleavage Activity 13-4
13.4. Example results 13-5
13.5. Conclusions 13-6

14. LOW-VOLUME, NON-TOXIC AND RAPID EXTRACTION OF SINGLE-STRAND-SPECIFIC NUCLEASES FROM CELERY 14-1

14.1 Objective 14-1
14.1. Materials 14-1
14.2. Methods 14-1
14.2.1. CEL I preparation 14-1
14.2.2. Activity tests 14-5
14.3. Conclusions 14-7
14.4. Contributors 14-7

15. MULTIVARIATE ANALYSIS – PHYLOGENETICS AND PRINCIPAL COMPONENT ANALYSIS 15-1

15.1. Phylogenetics 15-1
15.2. Inferring phylogeny from pairwise distances: construction of a distance tree using clustering with the unweighted pair group method with arithmetic mean (UPGMA) 15-2

15.3. Distance measures 15-2
15.4. Some reflexions on the comparison between genetic distances. 15-8
15.5. What genetic distance estimator to choose for essential derivation? 15-8
15.6. Genetic distances between populations 15-9
15.7. Protocol: tree reconstruction 15-10
15.8. UPGMA exercise 15-16
15.9. Principal Component Analysis (PCA) 15-20
15.9.1. Considerations and references 15-22
15.10. References 15-25

16. POPULATION GENETICS 16-1

16.1. Reading and coding genetic data 16-1
16.1.1. Presence/absence coding of dominant data 16-1
16.1.2. Allele size coding for microsatellites 16-2
16.1.3. Categorical coding 16-4
16.1.4. Presence/absence coding of co-dominant data 16-4
16.1.5. Formatting dominant data as co-dominant 16-5
16.1.6. Notes of formatting diploid data with spread sheets 16-6
16.1.7. Transforming data types using software 16-6
16.1.8. The FSTAT data file 16-7
16.2. Genetic diversity 16-8
16.3. Genetic structure 16-11
16.3.1. Nei's population genetics parameters: G_{st} family 16-11
16.3.2. Sewall Wright's F-statistics 16-11
16.4. Population and individual divergence and phylogenetic trees 16-12
16.5. Web resources and software – non-exhaustive 16-13
16.6. References 16-17
16.7. Some key concepts 16-19
16.8. Equations 16-1

17. APPENDICES 17-1

17.1. General DNA extraction techniques 17-1

17.1.1. Phenol/chloroform extraction 17-1
17.1.2. Ethanol precipitation 17-1
17.1.3. Solutions 17-1
17.2. Polymerase chain reaction protocol 17-2
17.2.1. References 17-6
17.3. Plant genome database contact information 17-7
17.4. Acronyms of chemicals and buffers 17-8

1. INTRODUCTION TO MOLECULAR MARKERS

Traditionally, molecular markers have played a major role in the genetic characterization and improvement of many crop species. They have also contributed to, and greatly expanded, our abilities to assess biodiversity, reconstruct accurate phylogenetic relationships, and understand the structure, evolution and interaction of plant and microbial populations. Molecular markers systems reveal variation in genomic DNA sequence and allow the tracking of this variation, ideally linked to phenotypic trait variation, in crossing programmes. The first generation of molecular markers, RFLPs, were based on DNA-DNA hybridisation and were slow and expensive. The invention of the polymerase chain reaction (PCR) to amplify short segments of DNA gave rise to a second generation of faster and less expensive PCR-based markers, which became popular in genotyping of many species. Today, next generation sequencing technologies have become the dominant tool for marker assisted breeding in developed countries and biotechnology companies. While incredibly powerful, these techniques are still cost-limiting and carry a heavy bioinformatics load, making use difficult in developing countries. This will likely change in the future as sequencing technologies and analysis tools increase in power and decrease in cost. Until then, we provide in this manual a series of low cost marker systems that are applicable in many laboratories with infrastructure for basic molecular biology.

Molecular markers are being used extensively to investigate the genetic basis of agronomic traits and to facilitate the transfer and accumulation of desirable traits between breeding lines. They are used both to tag target genes and to monitor the genetic background. A number of techniques have been particularly useful for genetic analysis. For example, collections of RFLP probes have been very versatile and important for the generation of genetic maps, construction of physical maps, the establishment of syntenic relationships between genomes, and marker assisted breeding. Numerous examples of specific genes that have been identified as tightly linked to RFLP markers are available for the improvement of specific agronomic traits in almost all major crops. Specific examples include viral, fungal and bacterial resistance genes in maize, wheat, barley, rice, tomatoes and potatoes. Additional examples include insect resistance genes in maize, wheat and rice as well as drought and salt tolerance in sorghum. These markers often used in conjunction with bulked segregant analysis and detailed genetic maps, provide a very efficient method of characterizing and locating natural and induced mutated alleles at genes controlling interesting agricultural traits. Markers have also been used to identify the genes underlying quantitative variation for height, maturity, disease resistance and yield in virtually all major crops. In particular, the PCR-based techniques have been useful in the assessment of biodiversity, the study of plant and pathogen populations and their interactions; and identification of plant varieties and cultivars. Amplified DNA techniques have produced sequence-tagged sites that serve as landmarks for genetic and physical mapping. It is envisioned that emerging oligonucleotide-based technologies derived from the use of hybridization arrays, the so-called DNA chips and oligonucleotide arrays, will become important in future genomic studies. However, many of these are still under development, are proprietary, or require the use of expensive equipment, and are therefore not yet suitable or cost-effective for adequate transfer to developing countries. Clearly, the initial transfer of technology has only involved a selected group of techniques that are well established and/or seem to have a broad application (*e.g.*, RFLP,

SSR, ISSR, AFLP, RAPD, IRAP and REMAP and SNPs). However, techniques are continuously changing and evolving, so technology transfer needs to keep pace with current developments in genomics. Capacity for handling molecular marker data has been identified as a bottleneck to the integration of molecular techniques in germplasm management. A module on population genetics, dealing specifically with the analysis of molecular marker data is included in this edition of the manual.

1.1. Use of molecular markers: A cautionary tale

Molecular biology is an exciting discipline with new techniques constantly being developed and high impact publications coming from the work. As such, it is tempting for the junior scientists to think of molecular tools as a starting point for their breeding objectives. The downside, however, is that these tools are often challenging to master, expensive and easy to mis-apply. It is important that experiments are carefully designed with proper controls and that the researcher understands the strengths and limitations of the chosen application. In this section we focus on the use of molecular markers. These tools can provide rapid, valuable information on the nucleotide diversity of collections allowing deductions of evolutionary relationships and gene flow. However, this manual is focused on *mutant* germplasm characterization, and when applying these tools for evaluation of induced mutant populations, an understanding of the genetics of the species and heritability of variation is required for proper application. To highlight this, we offer two different examples of application of markers; one correct, the other incorrect. If you are uncertain if molecular markers are right for you, please feel free to contact the Plant Breeding and Genetics Laboratory for further advice.

1.1.1. An example of how not to use molecular markers.

A research group is starting a new project to use induced mutations to breed for improved disease resistance in barley. They have never used induced mutations before and would like to use molecular markers to track disease resistance because it is very time consuming and expensive for them to test their material phenotypically at every generation.

The group produces a large M1 population that was treated with gamma rays. They self-fertilize the barley and grow the M2 in the next generation. They apply pathogen to the plants and score resistance. Of 10,000 plants, they find 50 with some increase in resistance to the pathogen. These 50 plants come from 20 different M1 parents. They collect tissue from these 50 plants, along with 10 mutagenized plants that are susceptible and 10 plants that were not mutagenized. They extract DNA, and perform an AFLP marker analysis. They hope to find bands that are common in the resistant plants but not in the control. Their data is not conclusive, so they decide to look at even more plants.

WHY IS THIS A BAD IDEA?

Current data suggests that most mutagenesis is random. In other words, different plants will have different changes in the DNA. Therefore, you don't expect the same mutations to be found in progeny from different M1 plants. Applying statistical probability, you might see this once or twice in a large population, but never 20 times. Therefore you don't expect to find bands in the mutants that arise due to common mutations.

BUT, HOW COME THEY ARE ALL DISEASE RESISTANT?

If a trait is polygenic, there may be many genes involved in a trait. Different plants in the example population may have mutations in different genes that give a similar phenotypic response. So, you don't need to mutate the same gene to get a similar phenotype. Additionally, there may be many possible mutations within the same gene that could give you a phenotype. The different alleles may not give the same signal in a marker assay.

1.1.2. An example of efficient application of markers

The researchers working with the barley population above have produced one line that is highly disease resistant after backcrossing to the parental line and applying selective pressure through five generations.

The issue with the parental line and the mutant line is that they are low yielding. The researchers would like to introgress the disease resistance into a high yielding cultivar that farmers are growing. To aid in this, the researchers apply a set of SSR markers to 300 plants from the disease resistance line, 300 parents and 300 of the elite variety. They identify one new band with a set of SSR primers that is present in all mutants but not in either the parent or the elite variety. They set out a crossing plan where they cross the mutant line with the elite variety. They self the F1s and then select only plants with the mutant SSR band. Starting in the F2, they select plants for disease resistance. They also apply AFLP and choose disease resistant plants that share the majority of markers with the elite variety.

WHY IS THIS A GOOD APPROACH?

The researchers have developed a marker by evaluating plants that are genetically related and harbouring the same mutation. Evaluation of a large number of plants allows the establishment that the marker is genetically linked to the mutation causing the phenotype. The lack of such bands in the control material reduces the risk that the marker is from some source of natural genetic variation. In the end, using AFLP allows for a high density of information on the genetic background of the selected individuals. It should be fairly straightforward to determine which plants have mostly elite variety background. This is what the breeder wants,

the elite variety with only that small amount of DNA conferring disease resistance introgressed, and not a lot of other DNA from the less suitable parent.

1.2. A Summary of Marker Techniques

Table 1.2–1. List of marker techniques

Marker/technique	PCR-based	Polymorphism (abundance)	Dominance
RFLP	No	Low-Medium	Co-dominant
RAPD	Yes	Medium-High	Dominant
SSR	Yes	High	Co-dominant
ISSR	Yes	High	Dominant
AFLP	Yes	High	Dominant
IRAP/REMAP	Yes	High	Co-dominant
Additional marker systems			
Morphological	No	Low	Dominant/Recessive/Co-dominant
Protein/isozyme	No	Low	Co-dominant
STS/EST	Yes	High	Co-dominant/Dominant
SNP	Yes	Extremely High	Co-dominant
SCARS/CAPS	Yes	High	Co-dominant
Microarray	Yes	High	

1.3. Ideal genetic markers

(highly dependent on application and species involved)

- No detrimental effect on phenotype
- Co-dominant in expression
- Single copy
- Economic to use
- Highly polymorphic
- Easily assayed
- Multi-functional
- Highly available (un-restricted use)
- Genome-specific in nature (especially when working with polyploids)
- Can be multiplexed
- Ability to be automated
- A perfect marker for the gene of interest, though for practical plant breeding a tightly linked marker is usually good enough.

1.4. Marker application suitability

RFLP	Comparative maps Framework maps, bin mapping Genetic maps Breeding Varietal/line identification (multiplexing of probes necessary) Marker-assisted selection F_1 identification Diversity studies Novel allele detections Gene tagging Bulk segregant analysis Map-based gene cloning
	This marker system is not suggested due to major issues in the lack of reproducibility.
SSR	Fingerprinting Varietal/line identification (multiplexing of primers necessary) Framework/region specific mapping Genetic maps F_1 identification Comparative mapping Breeding Bulk segregant analysis Diversity studies Novel allele detections Marker-assisted selection High-resolution mapping Seed testing Map-based gene cloning
ISSR	Fingerprinting Varietal/line identification Genetic maps F_1 identification Gene tagging Breeding Bulk segregant analysis Diversity studies Marker-assisted selection High-resolution mapping Seed testing
AFLP	Fingerprinting

	Very fast mapping Region-specific marker saturation Varietal identification Genetic maps F_1 identification Gene tagging Breeding Bulk segregant analysis Diversity studies Marker-assisted selection High-resolution mapping Map-based gene cloning
IRAP/REMAP	Fingerprinting Varietal identification F_1 identification Gene tagging Bulk segregant analysis Diversity studies Marker-assisted selection High-resolution mapping Seed testing
Morphological	Genetic maps Alien gene introduction Varietal/line identification F_1 identification Novel phenotypes Breeding
Protein and Isozyme	Genetic maps Quality trait mapping Varietal/line identification (multiplexing of proteins or isozymes necessary) F_1 identification Breeding Seed testing
STS/EST	Fingerprinting Varietal identification Genetic maps F_1 identification Gene tagging and identification Bulk segregant analysis Diversity studies Marker-assisted selection

	Novel allele detection
	High-resolution mapping
	Map-based cloning
SNP	Genetic maps
	F_1 identification
	Breeding
	Gene tagging
	Alien gene introduction
	Bulk segregant analysis
	Diversity studies
	Novel allele detections
	Marker-assisted selection
	High resolution mapping
SCARS/CAPS	Framework mapping
	Can be converted to allele-specific probes
	F_1 identification
	Gene tagging
	Bulk segregant analysis
	Diversity studies
	Marker-assisted selection
	Map-based cloning
Microarray	Fingerprinting
	Sequencing
	Transcription
	Varietal identification
	Genetic maps
	F_1 identification
	Gene tagging and identification
	Bulk segregant analysis
	Diversity studies
	Marker-assisted selection
	High-resolution mapping

1.5. Implementation

Table 1.5–1. Relative costs of marker techniques.

Marker/techniques	Development costs	Running costs per data point	Portability (Lab/Crops)
RFLP	Medium	High	High/High
RAPD	Low	Low	Low/Low
SSR	High	Medium	High/Low
ISSR	Low	Low	High/Low
AFLP	Medium-High	Low	High/Low
IRAP/REMAP	High	Medium	High/Low
Additional marker systems not covered in the course			
Morphological	Depends	Depends	Limited to breeding aims
Protein and isozyme	High	Medium	High/High
SCARS/CAPS	High	Medium	High/Low
STS/EST	High	Medium	Medium/High
SNP	High	Medium-Low	Unknown
Microarray	Medium	Low	Unknown

1.6. Requirements

Table 1.6–1. Requirements for marker techniques.

Marker/technique	Amount/ quality of DNA	DNA Sequence Required	Radioactive detection	Gel system
RFLP	High/High	No	Yes/No	Agarose
RAPD	Low/Low	No	No	Agarose
SSR	Low/Medium	Yes	No	Acrylamide/ Agarose
ISSR	Low/Medium	Yes/No	No	Acrylamide/ Agarose
AFLP	Low/High	No	Yes/No	Acrylamide
IRAP/REMAP	Low/Medium	Yes	No	Acrylamide/ Agarose
Additional marker systems not covered in the course				
Morphological	No	No	No	None

Protein/isozyme	No	No	No	Agarose/ Acrylamide
STS/EST	Low/High	Yes	Yes/No	Acrylamide/ Agarose
SNP	Low/High	Yes	No	Sequencing required
Microarray	Low/High	Yes	No	None
SCARS/CAPS	Low/High	Yes	Yes/No	Agarose

1.7. Comparison of different marker systems

Table 1.7–1. Advantages and disadvantages of various marker techniques.

Marker	Advantages	Disadvantages
RFLP	• Unlimited number of loci • Codominant • Many detection systems • Can be converted to SCARs • Robust in usage • Good use of probes from other species • Detects in related genomes • No sequence information required	• Labour intensive • Fairly expensive • Large quantity of DNA needed • Often very low levels of polymorphism • Can be slow (often long exposure times) • Needs considerable degree of skill • •
RAPD	• Results obtained quickly • Fairly cheap • No sequence information required • Relatively small DNA quantities required • High genomic abundance • Good polymorphism • Can be automated	• Highly sensitive to laboratory changes • Low reproducibility within and between laboratories • Cannot be used across populations nor across species • Often see multiple loci • Dominant
SSR	• Fast • Highly polymorphic • Robust	• High developmental and start-up costs • Species-specific • Sometimes difficult interpretation because of stuttering

Marker	Advantages	Disadvantages
	• Can be automated • Only very small DNA • Codominant • Multiallelic • Multiplexing possible • Does not require radioactivity	• Usually single loci even in polyploids
ISSR	• Highly polymorphic • Robust in usage • Can be automated	• Usually dominant • Species-specific
AFLP	• Small DNA quantities required • No sequence information required • Can be automated • Can be adapted for different uses, e.g. cDNA-AFLP	• Evaluation of up to 100 loci • Marker clustering • Dominant • Technique is patented • Can be technically challenging
IRAP/ REMAP	• Highly polymorphic depends on the transposon • Robust in usage • Can be automated • Species-specific	• Alleles cannot be detected • Can be technically challenging
Additional marker systems		
Morphological	• Usually fast • Usually cheap	• Few in number • Often not compatible with breeding aims • Need to know the genetics
Protein and Isozyme	• Fairly cheap • Fairly fast analysis • Protocol for any species • Codominant • No sequence information required	• Often rare • Often different protocol for each locus • Labour intensive • Sometimes difficult to interpret

Marker	Advantages	Disadvantages
STS/EST	• Fast • cDNA sequences • Non-radioactive • Small DNA quantities required • Highly reliable • Usually single-specific • Can be automated	• Sequence information required • Substantially decreased levels of polymorphism
SNP	• Robust in usage • Polymorphism are identifiable • Different detection methods available • Suitable for high throughput • Can be automated	• Very high development costs • Requires sequence information • Can be technically challenging
SCARS/CAPS	• Codominant • Small DNA quantities required • Highly reliable • Usually single locus • Species-specific	• Very labour intensive
Microarray	• Single base changes • Highly abundant • Highly polymorphic • Codominant • Small DNA quantities required • Highly reliable • Usually single locus • Species-specific • Suitable for high throughput • No gel system • Can be automated	• Very high development and start-up costs • Portability unknown

2. LOW COST DNA EXTRACTION WITHOUT TOXIC ORGANIC PHASE SEPARATION

One of the most common activities of molecular biology is the extraction of genomic DNA from cells. Traditional methods utilized lysis followed by organic phase separation to remove unwanted molecules such as proteins. Commercialized kits from companies such as Qiagen have circumvented unwanted toxic organic phase separation by using methods that employ DNA binding to silica with the use of chaotropic salts. This approach has proven superior in terms of speed and quality of product and has become the industry standard. The main issue with these commercial kits is that costs can become prohibitively expensive for large scale applications. The protocol below describes a home-made silica DNA binding protocol that costs about 1/10th that of a commercial kit and produces DNA quality suitable for TILLING and other high-throughput molecular applications.

2.1. Materials

	Company
MATERIALS FOR LOW-COST DNA EXTRACTIONS	
Celite 545 silica powder (Celite 545-AW reagent grade)	Supelco 20199-U
SDS (Sodium dodecyl sulfate) for mol biol approx 99%	Sigma L-4390-250G
Sodium acetate anhydrous	Sigma S-2889 (MW=82.03g/mol)
NaCl (Sodium chloride)	Sigma S-1314-1KG (MW=58.44g/mol)
RNase A	10 microgram per ml.
Ethanol	Ethanol absolute for analysis (Merck 1.00983.2500)
Nuclease-free H2O	Gibco ultrapure distilled water (DNase, RNase-free)
Guanidine thiocyanate	Sigma G9277 (MW=118.2g/mol)
Microcentrifuge tubes (1.5mL, 2.0mL)	Any general laboratory supplier
Micropipettes (1000µL, 200µL, 20µL)	Any general laboratory supplier
Microcentrifuge	Eppendorf Centrifuge 5415D
Optional: Shaker for tubes	Eppendorf Thermomixer comfort for 1.5mL tubes
MATERIALS FOR GRINDING OF LEAF MATERIAL (depending on grinding method)	
Liquid nitrogen	
Mortar and pestle or, TissueLyser, …	e.g. Qiagen TissueLyser II

Metal beads (tungsten carbide beads, 3mm)	Qiagen Cat.No. 69997 (for TissueLyser)
EVALUATION OF DNA YIELD AND QUALITY	
DNA concentration	ND-NanoDrop 1000 Spectrophotometer (optional)
Agarose gel equipment	Any supplier providing horizontal mini-gels
TILLING-PCR	
Thermocycler	Biorad C1000 Thermal cycler, or equivalent
PCR tubes	Life Science No 781340
TaKaRa Ex Taq™ Polymerase (5U/ul)	TaKaRA
10X Ex Taq™ Reaction Buffer	TaKaRa
dNTP Mixture (2.5mM of each dNTP)	TaKaRa
Agarose gel equipment	Any supplier providing horizontal mini-gels

2.2. Solutions to Prepare

BUFFER	Receipt		Comments
STOCK SOLUTIONS			
5M NaCl stock solution	MW=58.44g/mol 29.22g / 100mL		If keeping stocks for a long period, check to make sure high molarity stocks stay in solution. If precipitate forms, warm solution until back in solution, or discard and make fresh.
3M Sodium acetate (pH = 5.2)	MW=82.03g/mol 24.61g / 100mL		Adjust pH value with glacial acetic acid
95% (v/v) Ethanol	95 mL ethanol abs + 5 mL H_2O		
Tris-EDTA (TE) buffer (10x)	Composition: 100mM Tris-Cl, pH8.0 10 mM EDTA	For 100mL (10x): 10mL of 1M Tris-Cl stock 2mL of 0.5M EDTA stock	Tris and EDTA can be prepared from powder. Note that the pH of tris changes with temperature.
LYSIS BUFFER (standard)	0.5% SDS (w/v) in 10x TE 0.5g SDS /100mL		PBGL has developed a range of lysis buffers for different crops. If performance is poor, contact PBGL for modified buffers.
DNA BINDING BUFFER	6M Guanidine thiocyanate MW = 118.2 g/mol 70.92 g / 100mL (6M)		!!! it takes several hours until dissolved (leave it approx. 4-5 hours)
WASH BUFFER	1mL of 5M NaCl + 99mL of 95% EtOH		!!! PREPARE FRESH, because the salt precipitates during storage
DNA ELUTION BUFFER	depending on application (e.g. TE-buffer; Tris-HCl buffer)		

2.3. Methods (for centrifuge tubes)

PREPARATION OF SILICA POWDER-DNA BINDING-SOLUTION

- Fill silica powder (Celite 545 silica) into 50 mL-Falcon-tube (to about 2.5-mL = approx. 800mg)
- Add 30 mL dH_2O
- Shake vigorously (vortex and invert)
- Let slurry settle for approx. 15 min
- Remove (pipette off) the liquid
- Repeat 2 times (a total of 3 washes)
- After last washing step: resuspend the silica powder in about the same amount of water
  (up to about 5 mL)
- STORE the silica solution at RT until further use (silica : H_2O = 1 : 1)

Before use:

- suspend stored silica solution (silica : H2O = 1 : 1) by vortexing
- Transfer ~50 µL of silica solution to 2mL-tubes (prepare 1 tube per sample)
  NB try to keep the silica suspended during pipetting to ensure an equal distribution
- Add 1mL H_2O (a final wash step)
- Mix by vortexing
- Centrifuge: full speed (13.200) for 10-20 sec
- Pipette off liquid
- Add 700 µL DNA binding buffer (6M Guanidine thiocyanate)
- Suspend the silica powder in DNA binding buffer
- The silica binding solution is now ready for further use in the protocol (see Methods)

PREPARATIONS

- For TissueLyser: Prepare 2 mL-tubes (1 per sample): add 3 metal beads (tungsten carbide beads, 3mm) per tube
- Harvest leaf material (starting amount of material: about 100 mg fresh weight)

GRINDING

Use appropriate / available grinding protocol (mortar & pestle, Qiagen TissueLyser,)

For the TissueLyser:

- Freeze 2-mL tubes containing leaf material and 3 metal beads in liquid nitrogen
- Grind in TissueLyser by shaking (10 sec at 1/30 speed)
- Re-freeze in liquid nitrogen (>30 sec)
- Grind again in TissueLyser by shaking (10 sec at1/ 30 speed)
- Re-freeze in liquid nitrogen (>30 sec)
- Store in liquid nitrogen until lysis buffer is added

LYSIS

- Add 800µ **Lysis buffer**
- Add 4 µL **RNaseA** (10 µg/ml)
- Vortex (~2 min until the powder is dissolved in the buffer)
- Incubate: 10min at room temperature
- Add 200 µL **3M Sodium Acetate (pH 5.2)**
- Mix by inversion of tubes
- Incubate on ice for 5 min
- Centrifuge 13,200 rpm / 5 min / RT (pellet the leaf material)

DNA BINDING

- prepare **700 µL silica binding solution** (see above)
- transfer 800 µL of the supernatant to the tubes containing silica binding solution)

  !! Do not transfer leaf material!
- Completely resuspend the silica powder by vortexing and inversion of tubes (approx. 20 sec)
- incubate 15 min at RT (on a shaker at 400 rpm and/or invert tubes from time to time)
- Centrifuge 13,200 rpm / 3 min / RT (pellet the silica)
- Remove the supernatant (with pipette)

WASHING (2 times washing)

- Add 500 mL **wash buffer**

  !! Prepared fresh (see above)!
- Completely resuspend the silica powder by vortexing and inversion of tubes (approx. 20 sec)
- Centrifuge 13,200 rpm / 3 min / RT (pellet the silica)
- Repeat the washing step (optional: a third washing step)
- Remove the supernatant with pipette (as complete as possible)
- optional: short spin and remove residual liquid
- After last washing step: dry the silica in the hood up to 1 hour at RT (make sure there is no wash buffer left)

RESUSPENSION

- Add 200uL **TE buffer** or **10mM Tris buffer**
- Completely resuspend the silica powder by vortexing and inversion of tubes (approx. 20 sec)
- Incubate: 20 min / RT / with gentle agitation (on a shaker at 400 rpm and/or invert tubes from time to time)
- Centrifuge (for tubes): 13,200 rpm / 5 min / RT (pellet the silica)
- transfer 180 µL supernatant to new tube (avoid transferring silica powder!)
- optional: if there is still silica powder in the preps – repeat the centrifugation
- check for concentration and integrity of DNA
- store the genomic DNA at -20°C for long-term storage or 4°C for short-term storage

VALIDATION OF LOW-COST DNA PREPARATIONS FOR TILLING APPROACHES

Follow the protocol contained in "Positive control for mutation discovery using agarose gels, version 2.4"available at http://mvgs.iaea.org/LaboratoryProtocols.aspx , to test that your DNA is suitable for TILLING and Ecotilling applications.

2.4. Example Data

Table 1. Different combinations of self-made (low-cost) buffers and buffers from Qiagen DNeasy Plant Mini kit tested with barley tissue

Sample	1		2		3		4		5		6		7		8	
	+	-	+	-	+	-	+	-	A	B	A	B	A	B	A	B
Lysis	Dneasy kit* +Shredder columns -Shredder columns		Dneasy kit* +Shredder columns -Shredder columns		Dneasy kit* +Shredder columns -Shredder columns		Dneasy kit* +Shredder columns -Shredder columns		**Lysis buffer (PBGL)**		**Lysis buffer (PBGL)**		**Lysis buffer (PBGL)**		**Lysis buffer (PBGL)**	
DNA binding buffer	Buffer AP3/E*		Buffer AP3/E*		**6M Guanidine thiocyanate**		**6M Guanidine thiocyanate**		Buffer AP3/E*		Buffer AP3/E*		**6M Guanidine thiocyanate**		**6M Guanidine thiocyanate**	
DNA wash buffer	Buffer AW*		**Wash buffer – PBGL**		Buffer AW*		**Wash buffer-PBGL**		Buffer AW*		**Wash buffer-PBGL**		Buffer AW*		**Wash buffer-PBGL**	
DNA concentration (ng/µL)	14	13	34	41	7	8	4	10	12	11	12	20	10	16	13	17
Total yield (µg)	2.6	2.4	6.2	7.3	1.3	1.5	0.7	1.9	2.2	2.0	2.2	3.5	1.8	2.8	2.4	3.0
260/280 value	1.95	1.83	1.8	1.91	1.37	1.52	1.41	1.73	1.66	1.63	1.64	1.83	1.75	1.55	1.76	1.7

*components of Qiagen DNeasy Plant Mini kit

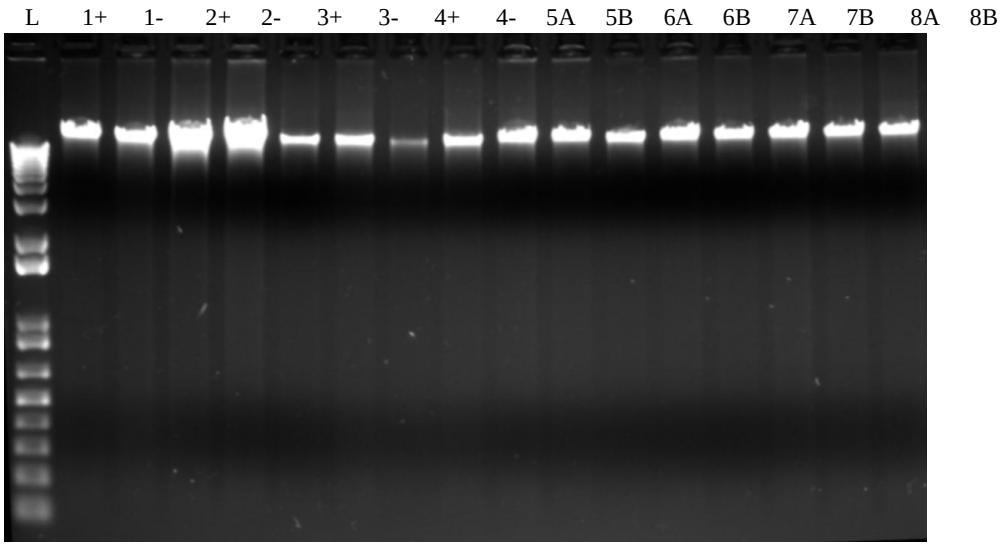


Figure 1. Quality of barley genomic DNA extractions using silica powder and different combinations of self-made (low-cost) buffers and buffers provided by Qiagen DNeasy kit. 8 µL of each genomic DNA extraction were separated on a 0.7% agarose gel.
1-8: Barley genomic DNA preparation
+: using QIAshredder columns for the preparation of barley leaf lysates (lysis procedure following the kit instructions)
-: preparation of leaf lysates using the kit instruction (but without using QIAshredder columns
A, B: technical replicates
L: size standard (1 kB Plus DNA ladder - Invitrogen)

All of the genomic DNA preparations show similar DNA concentrations (Table 1) and a good quality of the genomic DNA on the agarose gel (Figure 1). Only the DNA preparations "2+" and "2-" (buffer components from the kit in combination with our wash buffer) show clearly higher concentrations and yields (about 2-3 times higher) than all other DNA preparations. These results indicate that by modifications of the protocol (i.e. modifications of buffers) some improvements of the DNA yields are possible.
The DNA preparations of samples 8A and 8B were extracted exclusively with self-made (low-cost) buffers and show a comparable concentration and yield as the other extractions.

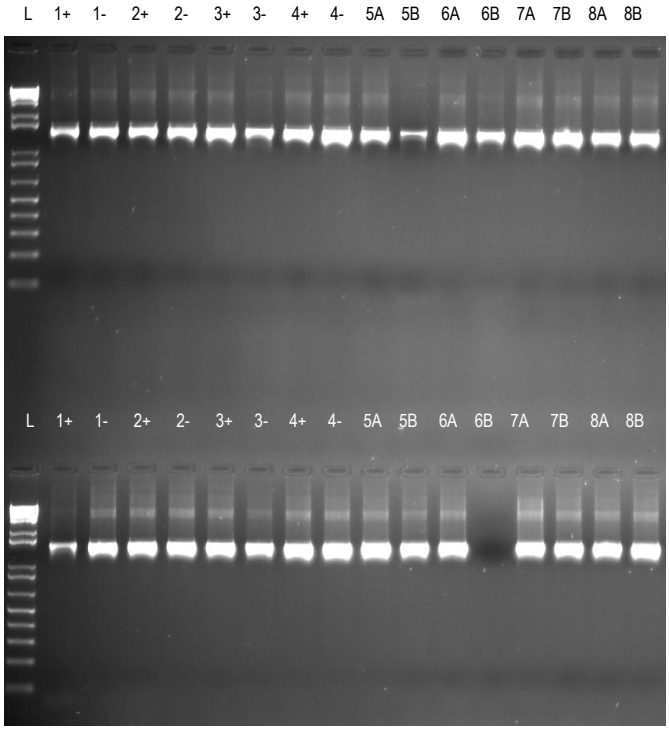


Figure 2. TILLING-PCR products amplified from genomic DNA extractions of barley (obtained by silica-based, low-cost DNA isolation method using different combinations of self-made buffers and buffers provided by Qiagen DNeasy kit). An aliquot of 5uL of each PCR reaction was separated on a 1.5% agarose gel.

top half – Target gene: nb2-rdg2a (1500bp-PCR product);
bottom half – Target gene: nbs3-rdg2a (1491bp-PCR product)
1-8: Barley genomic DNA preparation (see Table 1)
+: using QIAshredder columns for the preparation of barley leaf lysates – Lysis procedure following the kit instructions;
-: preparation of leaf lysates using the kit instruction (but without using QIAshredder columns
A, B: technical replicates
L: size standard (1 kB Plus DNA ladder - Invitrogen)

2.5. Conclusions

The DNA extractions from barley using the silica-based, low-cost method provided high-quality genomic DNA and sufficient yield suitable for standard PCR application such as molecular markers and TILLING.

3. DNA QUANTIFICATION

This protocol is designed to provide a standardized method for evaluating the quality and quantity of genomic DNA samples extracted from different plant species. Proper quantification and normalization of DNA samples to a common concentration is necessary prior to pooling samples for TILLING or Eco-tilling. A failure to combine genomes at an equal concentration can increase the false positive error rate because some polymorphisms will be represented at a concentration below the limits of detection.

3.1. Protocol for gel electrophoresis

3.1.1. Preparation of DNA concentration standards.

Lambda DNA (Invitrogen cat. # 25250-010) is used as a concentration standard.

A. Estimate how much concentration standard will be needed for a project (same organism, DNA prepared using the same methods, see 1.B.). Take this volume of DNA and vortex using the same settings as the genomic DNA extraction protocol used. This should shear the DNA to the approximate same size fragments as the genomic DNA. It is important to get the standard near to the same size as the genomic DNA because the intensity of ethidium bromide staining is a product of the size of DNA fragments.

B. Using the sheared DNA from 1.A, prepare DNA concentration standards at 115 ng/µl, 76.9 ng/ µl, 51.3 ng/ µl, 34.2 ng/ µl, 22.8 ng/ µl, 15.2 ng/ µl, 10.1ng/ µl, 6.8ng/ µl, 4.5 ng/ µl, and 3 ng/ µl. These are derived from the formula: 3×1.5^{i}, i = integers from 0 through 7. This is intended to provide the most accurate binning of DNA concentration estimates when performing visual analysis. Prepare the standards as independent dilutions from the stock of shaken Lambda to avoid cumulative error in low concentration DNA references. Prepare enough of each standard so that you have at least 3 µl for every 14 samples. Note that the concentration of lambda DNA may vary from batch to batch. Make sure to calculate dilutions based on the information printed on the stock tube.

3.1.2. Preparing agarose gels.

A. Prepare a 1.5% Agarose gel in 0.5x TBE buffer with 0.15 µg/ml ethidium bromide. Use at least a 24 tooth comb when preparing the gel. Place the solid gel into a rig containing 0.5x TBE buffer with 0.15 µg/ml ethidium bromide.

CAUTION: Ethidium bromide is mutagenic. Wear gloves, lab coat and goggle. Dispose of gloves in toxic trash when through. Avoid contaminating other lab items (equipment, phones, door handles, light switches) with ethidium bromide.

3.1.3. Preparing samples for loading into gels.

NOTE: When you have many samples to quantify, it is best to first test ~28 to determine the range of DNA concentrations from your extraction method. Samples above 62 ng/µl will be diluted to ~ 20 ng/µl for accurate quantification. If the majority of the small test subset have concentrations > 62 ng/ µl, you may want to dilute the rest of the samples prior to the agarose gel assay. This will save a gel run and the time required to estimate DNA concentrations.

A. Add 3 µl of DNA sample plus 2 µl DNA load dye (30% glycerol plus bromophenol blue – Do not add xylene cylanol as it migrates near the genomic fragment and can interfere with quantification). Use the same volumes for the DNA concentrations standards.

B. Load the gel. When using a 28 tooth comb, lanes 1-14 should contain genomic DNA samples and lanes 15-28 the concentration standard. Lane 15 should contain the 3 ng/ µl standard, lane 16 the 4.5 ng/ µl standard and so on with lane 28 containing the 115 ng/ µl standard.

3.1.4. Running the gel

A. Run gel at 5-6 V/cm (160V on a large Owl A2 rig, should be about the same for our rigs) for 30-60 min. The DNA sample should be completely out of the well and into the gel about 0.2 cm. Do not run the gel too long as the genomic DNA band will become diffuse and hard to quantify.

NOTE: Degraded samples (those producing smeary bands with standard agarose gels) should be run on a 3% MetaPhor agarose gel (~10.5g MetaPhor (Cambrex) in 350ml 0.5x TBE). The preparation of the MetaPhor gel is more specific in that it must be allowed to hydrate in the 0.5x TBE for ~15 min prior to melting. After melting and pouring, allow to set at room temperature, then put in the cold room (4°C) for 15-30 min. This final step is critical for proper setting of the gel.

3.1.5. Photographing the gel

It is important to get a proper exposure of the gel that shows a difference in ethidium staining in the concentration ranges you are assaying. For example, if all of your samples are at 20 ng/ µl, you should be able to observe a noticeable difference in the 34.2 ng/ µl, 22.8 ng/ µl, and 15.2 ng/ µl concentration standards. Make sure this is clear on the gel printout.

A. Adjust the image so as to take the longest possible exposure that does not saturate the image of any of the samples being assayed. It is all right to saturate the image of a reference sample that has higher [DNA] than any of the samples being assayed. Save this image in TIFF format. Print this image.

B. It may not be possible to set the exposure such that all bands can be visualized without saturating the higher concentration samples. In such a case, a second exposure is required for the notebook, but not for the scoring protocol on the gel documentation system as the computer can score samples that may be difficult to see by eye. Adjust the exposure of the gel so as to allow for the visualization of the lowest [DNA] samples. This will cause the saturation of the images of the highest [DNA] samples. Save this image as a TIFF file. Print this image.

3.2. Quantification of DNA using image analysis software

DNA concentrations can be estimated manually by comparing band intensity to the intensity of DNA standards of known concentration. A computer programme that capable of measuring pixel density can provide a more accurate and objective estimation of DNA concentration. In this method a standard curve is created with the DNA concentration standards and sample concentrations are estimated using the standard curve. Many GelDoc systems provide software for automated or semi-automated determination of DNA concentration based on pixel density. We provide here an alternative that will work on any digital tiff image using free image analysis software and Microsoft excel. The method can thus be applied to most labs.

1. The free programme ImageJ (http://rsbweb.nih.gov/ij/), is a public domain program developed by Wayne Rasband of the National Institutes of Health, USA Download this onto your computer. Full documentation can be obtained from the website.

2. Open ImageJ

3. Open the tiff image to be analysed (File>Open). A demonstration image titled “Cassava_DNA_test2c.tif” can be found on (URL) for practice.

CAUTION: Do not use compressed file formats such as jpeg.

4. Straighten the image so that the lanes are parallel with the image dialog box (Image>Rotate>Arbitrarily). In the rotate dialog box, select preview, set the Grid Lines number to 30, and adjust the angle in degrees until the bands are in line with the grid lines. The Interpolate feature should be selected. Note that you can set negative degrees by placing a minus (-) sign before the angle degree number. You may have to use a decimal setting to get the lanes to line up. When finished, click OK.

5. Subtract background noise (Process>Subtract Background). Deselect "light background". It is important that you don't set the rolling ball radius too small. It should be no more than half the width of the box you draw for the band (see step 7).

6. Select the rectangle tool in the ImageJ toolkit dialog box.

7. Find the highest intensity band on the gel to be analysed and draw a box around it. Make sure that the box surrounds the entire signal but does not overlap on the signal from another band. Check the height (h) and width (w) values and make sure that the larger of the two values is not more than 2x the size of the rolling ball radius chosen in step 5.

TIP: Select the magnifying tool and make the gel image as large as reasonable.

8. Left click and hold the mouse over the box and move it so that it is positioned around lane 1.

CAUTION: The box should contain only signal from the lane to be measured. Failure to do so will lead to an inaccurate reading.

9. Measure the box by hitting the m key. A full screen table should appear with columns for sample #, Area, Mean, min and max values. Minimize the table so that you can again view the gel image.

10. Move the box to lane 2 and hit the m key.

CAUTION: Do not change the size of the box. You must measure the same volume of box for each lane. If you accidentally change the size of the box while measuring lanes, start over.

11. Continue to move the box and hit the m key until all the lanes in a gel tier are measured, including the standards.

12. Evaluate the table. Does every sample have the same area value? If not, you have changed the size of the box and you need to start over. Does the number of samples equal the number of lanes on the gel? If not, you either missed a lane or counted a lane more than once. If so, you need to start over.

13. When you are satisfied that the table is correct. Select the entire contents of the table (control A), copy and paste into the raw data section of the excel worksheet.

14. Copy the density (area) from the last 6 samples representing the standards of known concentration in the test image. Paste these data into the density column just below the raw data. The excel table for the test gel image is found on (URL)

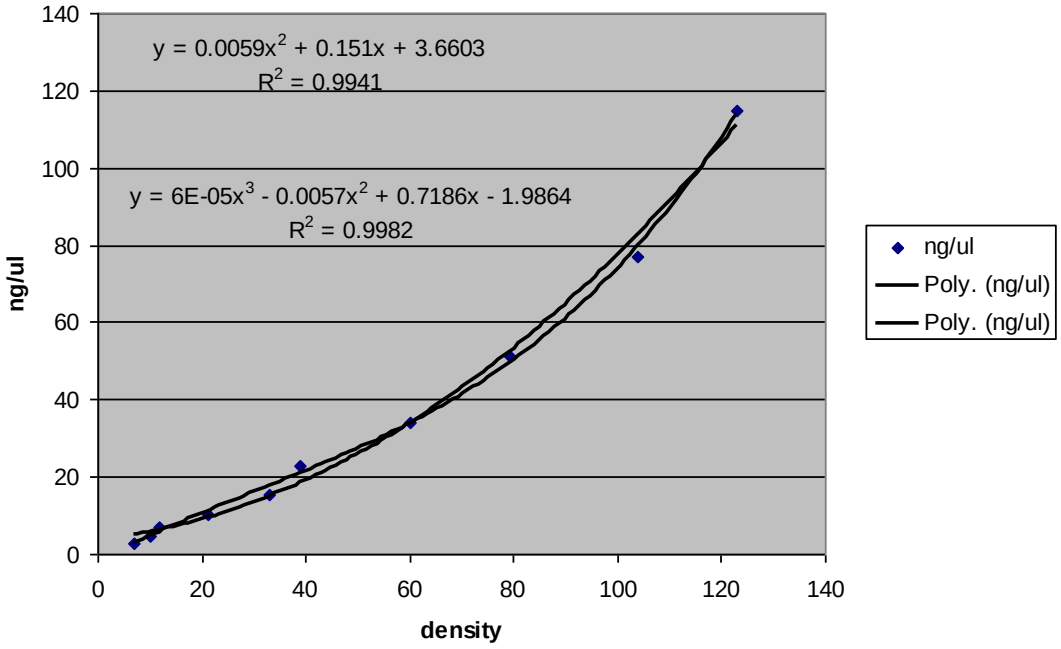
CAUTION: If you used less than the normal complement of standards, or put the standards in a different order than is represented in the "ng/µl" column, you will need to modify this section appropriately.

15. Select the density and ng/µl columns including the title cells (A, B 41-47 in excel). Click the "Chart Wizard" button. Select XY (Scatter) as chart type and scatter with no point connection as sub type. Click next

16. Select the series in columns. Click next and fill out the title (Gel #), X axis (density) and Y axis (ng/µl). Click next and save the graph as an object in the workbook. Click finish. Move this graph to the graph section of the worksheet.

17. Inspect the graph. Are there any points that are clearly off of the trend? If so, consider removing this data point and re-drawing the graph. This may become more evident once you have drawn the trendline (Step 18).

18. Add a trendline (Chart>Add Trendline). Under type, click polynomial and select 2nd order. Click Options and select "Display equation on chart, and display r-squared value on chart. Click ok. OPTIONAL: You may try a higher order polynomial to evaluate how differences in curve fitting can affect your concentration estimation (see figure below showing second and third order polynomial).



19. Fill in the sample # next to the lane number in the DNA concentration table to the right of the raw data section.

20. Copy and paste the density from the raw data into the density column of the concentration table to the right of the raw data.

21. Insert the formula for the second order polynomial into the first cell of the second order polynomial column. Copy the formula from the graph, then click on the cell, type the equal (=) symbol in the formula box and paste the formula. Replace x^2 with the density data from the first sample. This sample should be in cell J7, so you would replace x^2 with *j7*j7. Replace x with *j7. When finished, press the enter key. The value should appear in the cell.

22. Click on the cell. Pull the right corner so that the box extends over the entire column. You should see all the cells in that column fill with the appropriate values.

Optional. Repeat Step 21 and 22 for the third order polynomial. For x^3, use *j7*j7*j7. For many cases the second order polynomial will be sufficient. The main differences will be in estimating high (>50 ng/µl) concentration samples.

23. Save the gel image in imageJ as a tif image in a new folder labelled with the gel image name.

24. In the excel workbook, import the tif gel image and place it near the Gel Image field.

25. Compare the band intensities on the image with the concentrations estimated from the standard curve. Do you agree with the estimations? If not, consider repeating the measurement.

26. Compare your data with the data provided in the sample data tab of the excel sheet. Did you get the same results?

4. RESTRICTION ENZYME DIGEST

Restriction enzymes are produced by various bacterial strains. In these bacterial strains they are responsible for limiting attack from certain bacteriophages. They act by cutting ("restricting") the phage DNA at a sequence-specific point, thereby destroying phage activity. Sequence-specific cutting is a fundamental tool in molecular biology. DNA fragments can be ligated back together ("recombined") by T4 DNA ligase. In addition to cloning and molecular marker applications, restriction digestion is being used for new techniques such as for creation of restriction phased libraries for Next Generation Sequencing (NGS). Many restriction enzymes have been cloned and are available in a commercially pure form. They are named after their bacterial origin: e.g. *Eco*RI from *E. coli*.

The known restriction enzymes recognize four or six bases (eight in the case of "very rare cutters" like *Not*I and *Sfi*I). Recognition sequences are almost always "palindromic" where the first half of the sequence is reverse-complementary to the second:

e.g. the *Xba*I site is

5' T C T A G A 3'
3' A G A T C T 5'

The position of the actual cut is enzyme dependent and symmetrical on the opposite strand:

5' T C T A G A 3'
3' A G A T C T 5'

leaving cohesive termini (sticky ends) at the 5' end:

5' T 3'
3' A G A T C 5'

The commercially available restriction enzymes are supplied with the appropriate restriction buffers (10 x concentrated). The enzymes are adjusted to a specific activity per µl, usually 10 U/µl. (1 Unit is the amount of enzyme needed to cut 1 µg of lambda DNA in one hour at 37°C).

A typical restriction digestion is performed using between 20µl and 100µl reaction volume per 5 µg and more of plant DNA. For purified plasmid DNA 2 U per µg DNA is sufficient, for plant DNA 4 U per µg should be used.

For example: digestion of 5 µg DNA in 40 µl reaction volume:

Restriction buffer (10x)	4 µl
DNA 1 µg/µl	5 µl
Doubled distilled H_2O	29 µl
Enzyme (10 U/µl)	2 µl

Incubate for at least 1 hour at 37°C. The restriction enzyme can be inactivated by heating to 65°C for 10 minutes or by adding 1.0µl 0.5 M EDTA.

Note however, that protein engineering and advanced biochemistry have allowed major improvements from the canonical restriction digestions above. For example, Thermoscientific have developed a suite of fast enzymes that can digest complete genomes in 15 minutes, versus the traditional overnight digestion. Such digestions can be accomplished with no star activity.

5. FINDING CANDIDATE GENES AND PRIMER DESIGN FOR MOLECULAR TESTING: AN EXAMPLE FROM THE ANNOTATED *SORGHUM BICOLOR* GENOME.

5.1. Overview

There are several levels of genome annotation. The goal of this method is to quickly identify annotated genes and recover gene and transcript/protein sequences from the Sorghum genome that have potentially interesting biological function, without extensive bioinformatics expertise or tools. The same methods can be applied to many other annotated genomes. Genome project websites typically have text files of genome annotations. Many genome projects use the same generic genome browser architecture, and so retrieval of sequences described here will work for different species. For example, there are many genomes available on Phytozome.

Retrieve a list of annotated genes in the Sorghum genome.
This file: ftp://ftp.jgi-psf.org/pub/compgen/phytozome/v8.0/Sbicolor/annotation/Sbicolor_79_annotation_info.txt

while not the most verbose annotation it is easily opened and searchable.

Open this file up and hit control F, you can do a quick text match search for keywords like disease. If you search for disease, you get >100 hits.
The first hit for a text search of disease is Sb0019s003010.1.

Recover sequences for your favourite gene
There are (at least) two ways to retrieve the sequence for primer design.

First, you can search NCBI (http://www.ncbi.nlm.nih.gov). You need to remove the ".1" at the end because this delineation is not in NCBI. What you'll get is an 800,000 bp scaffold that contains the gene sequence. Unfortunately, it contains many predicted proteins, but the annotation isn't there. Which means that it is very hard to find the protein you're looking for unless you blast all the hypothetical peptides. This isn't very convenient.

To retrieve genomic, cDNA and protein sequences, goto the genome website http://www.phytozome.net/sorghum. Click "Browse Genome" and then enter Sb0019s003010.1 into the landmark or region window and click search. You'll get the gene model back with blast hits to other plant proteins. Move the mouse over this pile up and you'll get individual annotations from the different species (this is good to do to double check you have the correct gene).
4. Download sequences for downstream analysis and primer design.

In many cases (such as TILLING/Ecotilling) it is best to be searching for potentially functional variation. So, it will be more efficient to screen exonic regions. In this example notice the exonic regions are mostly on the left side.
It is not very intuitive how to get bot the genomic and transcript sequence from this graphical output. Put your mouse over the transcript and right click. A new window will appear from phytozome and you can get the sequences you need from the sequencing tab.

FOR TILLING and Ecotilling applications design primers following protocol in chapter section 13.2.1.

6. RFLP

RFLP definition: The variation(s) in the length of DNA fragments produced by a specific restriction endonuclease from genomic DNAs of two or more individuals of a species (Kahl, 2001).

Restriction fragment length polymorphism (RFLP) technology was first developed in the 1980s for use in human genetic applications and was later applied to plants. By digesting total DNA with specific restriction enzymes, an unlimited number of RFLPs can be generated. RFLPs are relatively small in size and are co-dominant in nature. If two individuals differ by as little as a single nucleotide in the restriction site, the restriction enzyme will cut the DNA of one but not the other. Restriction fragments of different lengths are thus generated. All RFLP markers are analysed using a common technique. However, the analysis requires a relatively complex technique that is time consuming and expensive. The hybridization results can be visualized by autoradiography (if the probes are radioactively labelled), or using chemiluminesence (if non-radioactive, enzyme-linked methods are used for probe labelling and detection). Any of the visualization techniques will give the same results. The visualization techniques used will depend on the laboratory conditions.

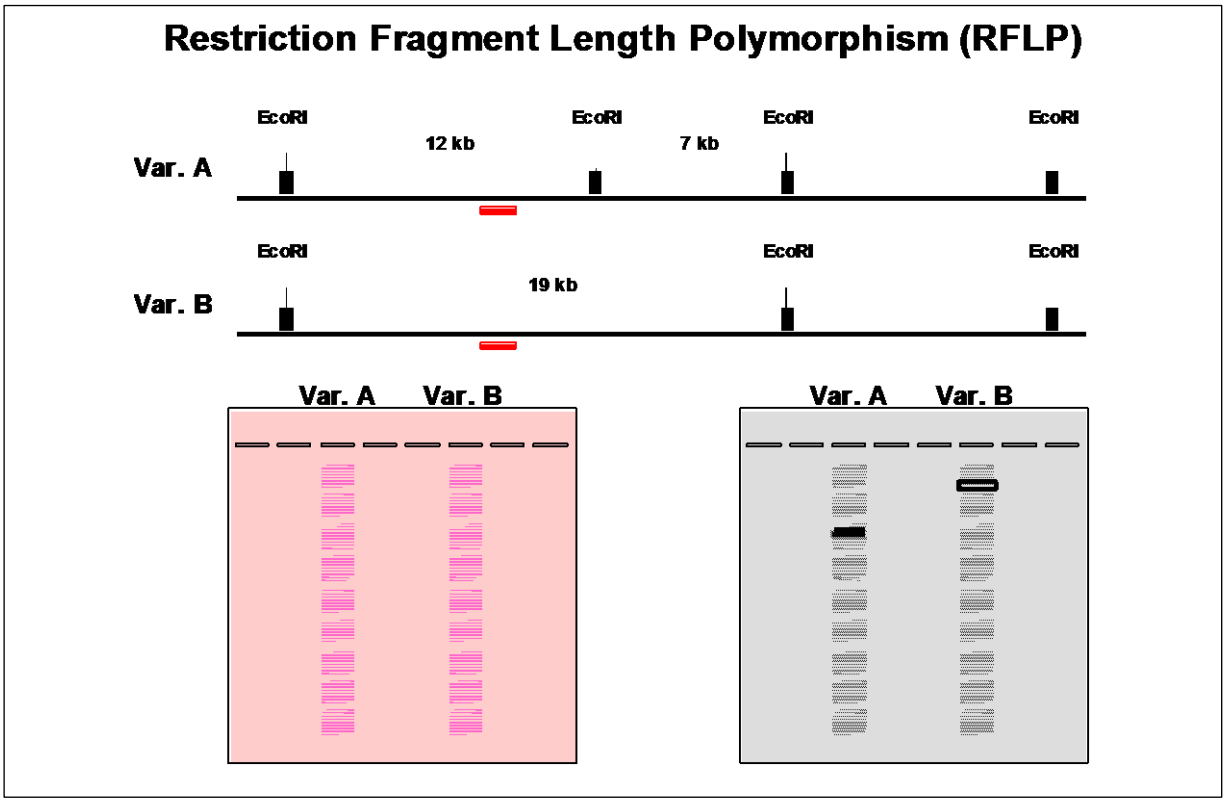


Figure 6-1. The scheme depicts enzyme digestion of DNA into fragments and their subsequent gel separation and the detection of allelic variation in varieties A and B (with permission, K. Devos).

Figure 6-2. An autoradiograph detecting parent (P1 and P2) and homozygous (1 and 2), respectively and heterozygous (H) F_2 segregation (with permission, M.D. Gale).

6.1. Protocol

6.1.1. Agarose gel electrophoresis

Agarose is a galactose-based polymer, widely used in analytical and preparative electrophoretic separation of linear nucleic acids in the size range above 100 bp. DNA applied to an agarose gel, which is exposed to an electrical field, migrates towards the anode, since nucleic acids are negatively charged. The smaller the molecules the faster they run through the gel matrix (Figures 5.1, 5.2, and 5.3). Migration is inversely proportional to the log of the fragment length. In order to determine the length of the separated fragments in the gel a molecular weight fragment ladder control is placed in a lane alongside the experimental samples. Restricted genomic DNA is usually separated in a 0.8-1.0 % gel whereas gels with a higher concentration of agarose (2-3%) are needed for separation of small DNA fragments (<500 bp).

Method: Gel preparation and running

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination of preparation.

NOTE: The buffer for gel preparation and for filling the electrophoresis tank is 0.5xTBE.

1. Agarose powder is dissolved in buffer by slowly boiling in a microwave or water bath.

2. Let the agarose cool down to 60°C (just cool enough to hold).

3. Ethidium bromide (EthBr) is added to the gel at a concentration of 0.5 µg/ml before the gel is poured (alternatively the gel can be stained after electrophoresis in water containing EthBr). (*Caution: Ethidium bromide is toxic. Gloves should be worn and avoid inhalation.*)

4. As the agarose is cooling down prepare the gel tray by placing tape across the ends of gel tray such that there is no leakage and so the tray will be able to accommodate the desired thickness of the gel.

5. Pour the agarose-EthBr mixture into the prepared gel tray and insert combs using a comb size depending on the depth, width, and thickness of the desired well. To avoid breaking the wells when the comb is removed, leave 1mm between the comb teeth and the bottom of the gel tray. Allow the gel to solidify (20-30 minutes).

6. Remove tape and place tray in gel rig. Pour enough 0.5x gel buffer into the gel rig to cover the gel, then remove combs.

7. Load the DNA samples, containing the lane marker bromophenol blue dye, into the wells. Load the wells of the gel to the top. It typically takes 30 to 40 µl to fill each well.

NOTE: Do not over load the wells as that would definitely lead to DNA contamination.

NOTE: The DNA is mixed with loading buffer and dye order to facilitate the solution sinking into the gel wells. As a single band, 10 ng DNA can still be visualized with EthBr.

8. Run samples into gel at 100mA for 5-10 minutes, then reduce the amperage and run at 25 mA, constant current, until the bromophenol blue dye marker has migrated almost to the end of the gel. Typically a long gel will be done after 14-16 hours.

NOTE: The following step is used only if the EthBr was not added as in step S.3. Stain each gel in 1 µg/ml EthBr (50 µl of 10 mg/ml EthBr in 500 ml dH_2O) for 20 minutes shaking gently.

9. Rinse gel in ddH_2O for 20 minutes, slide gel onto a UV transilluminator and photograph. For Fotodyne PCM-10 camera with 20 x 26 cm hood and Type 667 Polaroid film use an f8, 10 second exposure. (*Caution: Wear gloves and lab coat, and UV-protective full face shield or glasses when you are exposed to the UV light of the transilluminator.*)

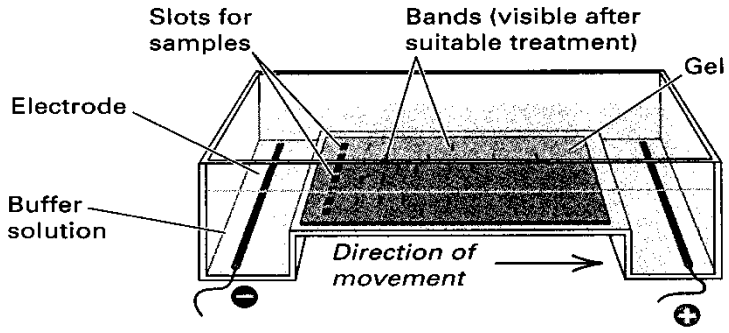


Figure 6-3. Apparatus for gel electrophoresis (Hartl and Jones, 1999).

6.1.2. Southern blotting and hybridization

Southern blotting

Localization of particular sequences within genomic DNA is usually accomplished by the transfer technique described by Southern (1975) and subsequent hybridization with a labelled probe. Genomic DNA is digested with one or more restriction enzymes, and the resulting fragments are separated according to size by electrophoresis through an agarose gel. The DNA is then denatured *in situ* and transferred from the gel to a nylon membrane. The relative positions of the DNA fragments are preserved during their transfer to the filter. The DNA is hybridized to radioactive or (in our case) non-radioactive labelled DNA probes, and the positions of bands complementary to the probe can be visualized by autoradiography or alternative enzyme-linked detection systems.

Capillary transfer: In the capillary transfer method (Southern 1975), DNA fragments are carried from the gel in a flow of liquid and deposited on the surface of the nylon membrane. The liquid is drawn through the gel by capillary action that is established and maintained by a stack of dry and absorbent paper towels (see Figure 5.4).

Method: Transfer of DNA from agarose gel to a nylon membrane.

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

1. After taking a photograph of the gel mark the gel for orientation purposes.
2. Soak the gel for 5 minutes in 0.25 M HCl for depurination.
3. Soak gel 2 x 20 minutes in denaturing solution (0.4 M NaOH, 1 M NaCl) with constant, *gentle* agitation. Meanwhile prepare the transfer apparatus (see Figure 5.4).
4. Discard denaturing solution and add 1M ammonium acetate to neutralize the gel (shake for 10 minutes).

5. Wrap a piece of Whatman 3 MM paper around a piece of Plexiglas or a stack of glass plates to form a support that is longer and wider than the gel (an empty box for pipette tips with plain surface is sufficient). Place the wrapped support inside a large baking dish, which is then filled with the transfer buffer (20XSSC).
6. Cut a piece of nylon membrane to the size of the gel along with a similar piece of 3 MM paper (do not touch the membrane, wear gloves and lab coat, and use forceps to handle membrane - otherwise it will result in background signals after detection). Wet both pieces in transfer buffer. Place the gel *face-down* on the wrapped support, and smooth out all bubbles.
7. Place the nylon membrane on top of the gel and smooth out all bubbles. Cut a corner of the membrane according to the orientation cut made on the gel. Mask the surrounding 3 MM paper with Parafilm strips.
8. Place the wet 3 MM piece on top of the membrane, excluding bubbles, followed by a further dry piece and then a stack of paper towels (5-8 cm high). Put a glass plate on top of the stack.
9. Wrap the whole apparatus with clingfilm to reduce evaporation and weigh the stack down with a 500 g weight.
10. Leave overnight for transfer - and sleep well!
11. Remove the paper towels and the 3MM paper from the gel. Peel the membrane off and soak it for 5 minutes in 2XSSC to remove any pieces of agarose sticking to the filter.
12. Dry the membrane on 3 MM paper for 30 minutes.
13. Then fix the DNA by baking the filter (refer to manual of the nylon membrane which is used, *e.g.* the positively charged nylon membrane from Roche is baked for 30 minutes at 120°C).
14. Proceed with hybridization of probe.

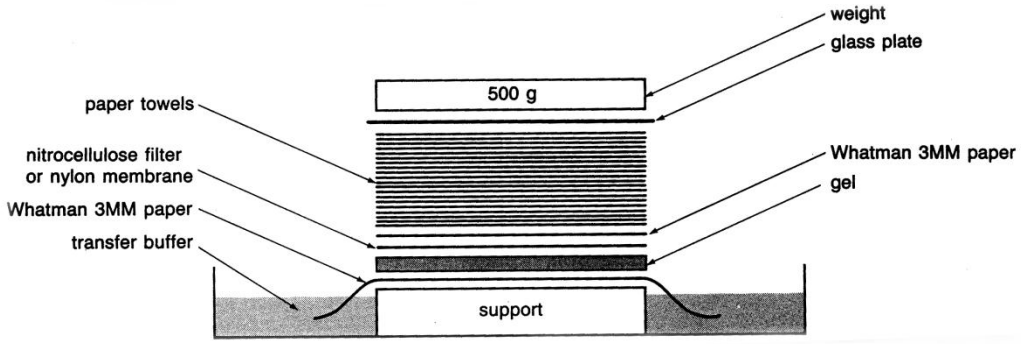


Figure 6-4. Blotting apparatus for capillary transfer of DNA (Sambrook *et al.*, 1989)~~(Sambrook *et al.*, 1989)~~.

DNA:DNA hybridisation using the DIG system

NOTE: The hybridization protocol used in the FAO/IAEA course was that obtained in the "Random Prime Labelling and Detection System" (RPN 3040/3041) commercially available from Amersham LIFE SCIENCE. This is a very good labelling and detection kit that comes

with a step-by-step procedure. However, if you cannot obtain the Amersham kit try the following protocol, which also works well.

Most of the *non-radioactive* labelling and detection systems for nucleic acids are based on the incorporation of a nucleotide, which is linked to a haptene molecule, into the hybridization probe. The identification of the haptene molecule at the hybridization sites is facilitated by an immunological detection reaction.

In the case of the Digoxigenin system (DIG-system, Roche) the haptene is digoxigenin, a steroid exclusively occurring in the plant *Digitalis purpurea*. The molecule is linked to desoxyuracilphosphate by an 11 atoms linear spacer (Dig-[11]-dUTP), (Figure 5.5).

The DNA:DNA hybridization sites are detected by using antibodies against digoxigenin, which are conjugated to alkaline phosphatase (AP) as a reporter enzyme. By adding the colourimetric substrate NBT/X-phosphate or alternatively the chemiluminescence substrate AMPPD (CSPD®) the presence of the enzyme is visualized (Figure 5.6).

The main advantage of the non-radioactive system is the avoidance of radioisotopes and the associated hazards, as well as saving high costs for maintaining an isotope laboratory (e.g. for disposal of the radioactive waste). Furthermore, DIG labelled probes are much more stable. They can be stored at -20°C for more than 12 months, and the hybridization solution can be re-used several times. At the same time the sensitivity of the DIG system is comparable to that of ^{32}P labelled probes.

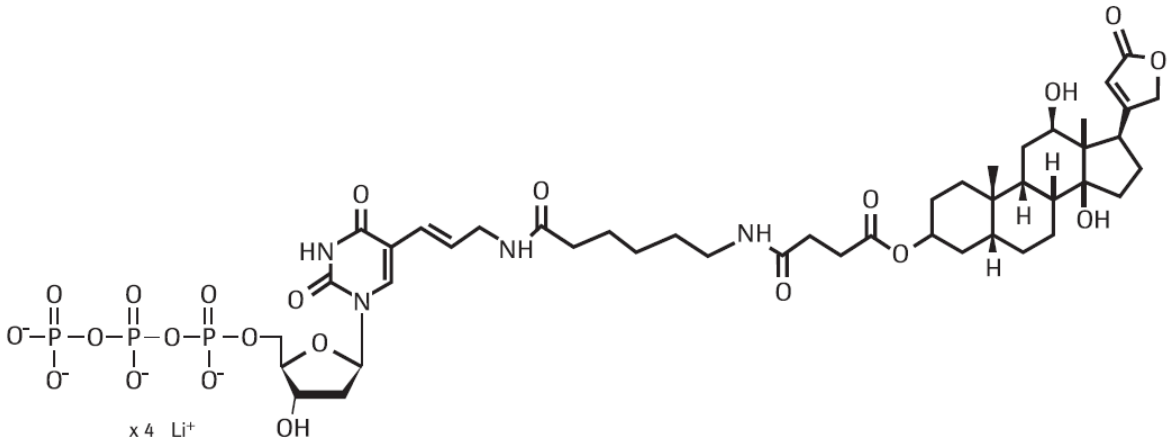


Figure 6-5. Structure of the Dig-[11]-dUTP molecule (source: DIG DNA Labelling and Detection Kit).

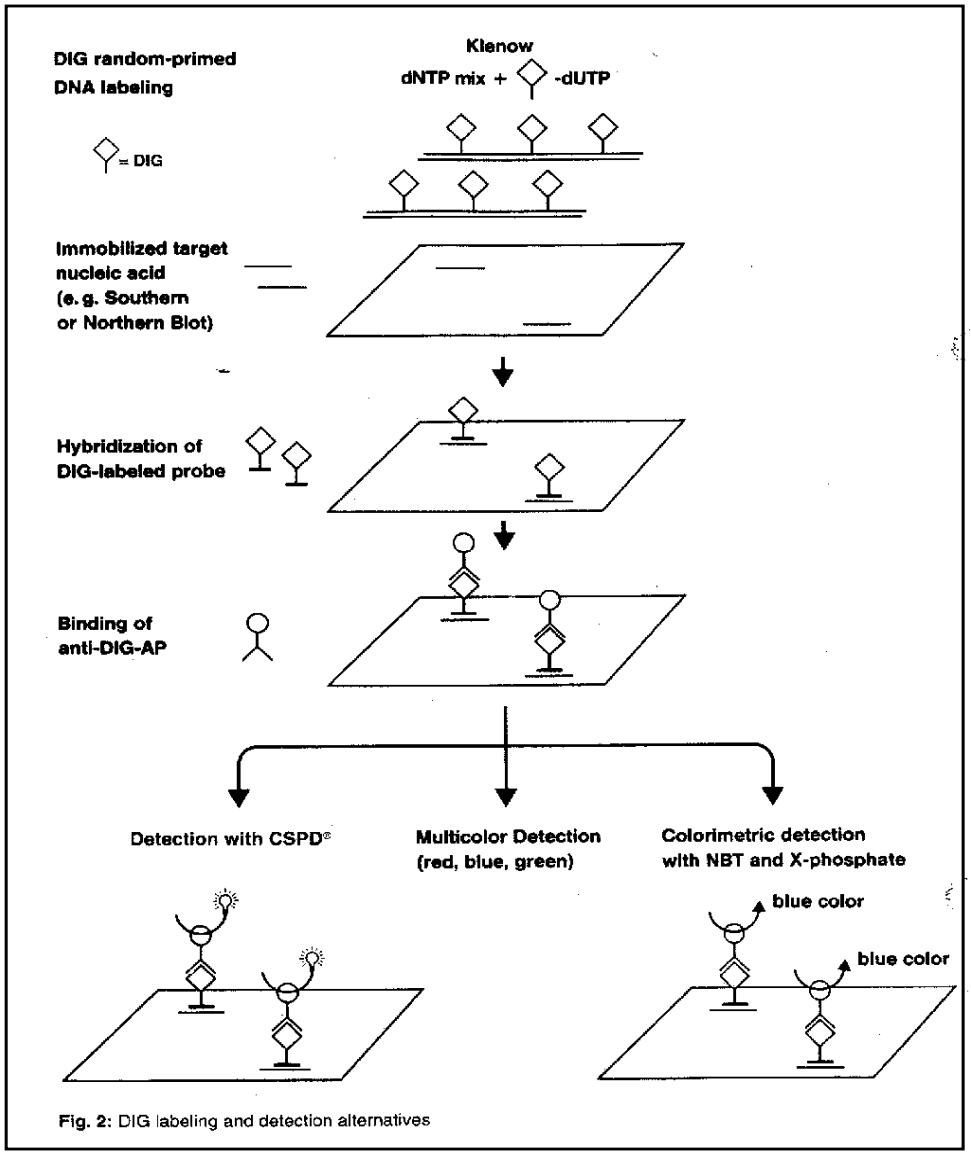

DIG random-primed DNA labeling
Klenow
dNTP mix + -dUTP
= DIG
Immobilized target nucleic acid (e. g. Southern or Northern Blot)
Hybridization of DIG-labeled probe
Binding of anti-DIG-AP
Detection with CSPD®
Multicolor Detection (red, blue, green)
Colorimetric detection with NBT and X-phosphate
blue color
blue color
Fig. 2: DIG labeling and detection alternatives

Figure 6-6. DIG labelling and detection alternatives (source: DIG DNA Labelling and Detection Kit).

6.1.3. Labelling the probe and dot blot/quantification

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

1. Dilute template DNA (0.5 µg - 3 µg) to a total volume of 15 µl and denature by heating for 10 minutes in a boiling waterbath, then quickly chill on ice/NaCl.
2. Add on ice: 2 µl hexanucleotide-mixture, 2 µl dNTP mixture (containing Dig-[11]-dUTP), and 1 µl Klenow enzyme (DNA polymerase).

3. Mix, centrifuge briefly, and then incubate for at least 60 minutes (20 hours is better) at 37°C.
4. Add 2 µl 0.2 M EDTA, pH 8.0 to stop the reaction.
5. Precipitate the labelled DNA by adding 2.5 µl 4M LiCl and 75 µl pre-chilled ethanol. Mix well and leave for 2 h at -20°C.
6. Spin in a microcentrifuge for 15 minutes. Wash the pellet with 50 µl cold ethanol, 70%.
7. Dry the DNA pellet and dissolve in 50 µl TE-buffer.
8. Dot Blot/Quantification of labelling efficiency

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

It is absolutely necessary to estimate the yield of DIG-labelled probe. If the probe concentration in the hybridization solution is too high, large background signals will appear on the blot after detection. Therefore the kit contains a DIG-labelled control DNA of known concentration. A dot blot with a dilution series of your probe and the provided control DNA makes the test. If the amount of template DNA was about 1,000 ng you can expect between 260 ng (after 1 hour incubation with Klenow enzyme) up to 780 ng (after 20 hours) of newly synthesized DIG-DNA.

8.1. Cut a piece of nylon membrane and label 1 cm^2 squares with a soft pencil.

NOTE: Do not use an ink or ballpoint pen.

8.2. Apply 1 µl of the probe dilution series (1:10, 1:100, 1:1,000) and of the control-DNA dilution series to each square on the membrane. To prepare the dilution series of the control DNA follow the scheme proposed in the kit manual (see below).

8.3. Fix the DNA to the membrane by cross-linking with UV-light or baking (dependent on the type of nylon membrane used).

8.4. After the spots are dry continue with the detection procedure. The colourimetric assay is the method of choice, because you can easily follow the development of the colour on the membrane.

8.5. Stop the reaction as long you can see differences between the concentrations of the calibration series. For detection procedure, see below (5.1.2.6).

Table 6.1–1.

DIG-labelled control DNA, diluted 1:5; starting concentration (1µg/ml)	Stepwise dilution in DNA dilution buffer	Final concentration (pg/µl)	Total dilution
(A) 1ng/µl	5µl/45µl	100	1:10
(B) 100pg/µl	5µl/45µl	10	1:100
(C) 10pg/µl	5µl/45µl	1	1:1,000
(D) 1pg/µl	5µl/45µl	0.1	1:10,000
(E) 0.1pg/µl	5µl/45µl	0.01	1:100,000

6.2. Hybridisation

Pre-hybridization

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination of your preparation.

1. The nylon membrane is inserted into a heat resistant polythene bag.
2. The hybridization solution (without the probe!) is added (20 ml per 100 cm^2 membrane).
3. Before heat-sealing the bag, air bubbles are removed by rolling a pipette over the bag, which should be placed on a sloping plane.
4. Allow the sealed bag to gently shake in the water bath at 42°C for at least 1 hour.

Hybridization

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

1. The pre-hybridization solution can be exchanged with the hybridization solution containing the probe (2.5 ml per 100 cm^2 membrane), or as an alternative, you could add the probe to the hybridization mixture directly into the bag.
2. The DIG-labelled probe has to be denatured as before (see Section 4.1.2.3 Step S.1) and is subsequently added to the hybridization solution at a concentration of 40 ng/ml (for probe concentration see results of dot blot test (see Section 5.1.2.3 Step S.8.5).
3. Carefully remove all air bubbles from the bag before you heat-seal it.
4. Let the hybridization proceed overnight (at least 14 hours) in the water bath at 42°C (with formamide-containing hybridization solution) or 68°C (without formamide) with gentle agitation. [*Caution: Formamide is harmful. Gloves should be worn.*]

NOTE: After hybridization the solution is collected at one corner of the bag by rolling a pipette over it and transferred to a reaction tube for re-use.

5. The hybridization solution containing the Dig-labelled probe is stored at -20°C and can be re-used several times.

NOTE: It has to be denatured before each new application.

6.2.1. Washing method

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

During the washing procedure the remaining probe is diluted and washed from the membrane. In a second washing step the probe DNA, which binds unspecifically to the DNA on the blot, is removed.

It is useful to know that the stability of DNA:DNA hybrids is dependent on certain factors, such as the melting temperature (T_m) at which the probe is annealed to 50% of its exact

complement. The factors influencing the T_m are included in the formula of Meinkoth and Wahl (1984):

T_m = 81.5°C + 16.6 log *M* + 0.41 (% G + C) - 500/*n* - 0.61 (% formamide) where *M* is the concentration (mol l^{-1}) of monovalent cations in the hybridization solution/washing solution, (% G + C) the proportion of guanine and cytosine in the probe, and *n* the length of the probe in base pairs.

The melting temperature T_m together with the selected hybridization and washing temperature T_a determine the conditions for annealing between probe and target DNA. This is called the stringency:

stringency (%) = 100 - M*f* (T_m - T_a)
where M*f* is the "mismatch factor" (1 for probes longer than 150 bp).

Under hybridisation/washing conditions with a stringency of 100%, all DNA:DNA hybrids with less than 100% homology are resolved.
In general one can say, the lower the salt concentration in the washing solution and the higher the hybridization or washing temperature, the higher the stringency.

1. The hybridization bag is opened.
2. The membrane is transferred to a plastic dish. It is very important that the plastic dish has been thoroughly cleaned. Use 500ml of each solution per 100cm^2 membrane.

   1st wash: 2 x SSC, 0.1% SDS (w/v) - 2 x 15 minutes at room temp.
   2nd wash (new dish): 0.1XSSC*****, 0.1% SDS (w/v) - 2 x 15 minutes at 68°C.

NOTE: *****These conditions are highly stringent. The SSC concentration in the second (stringent) wash should be increased when a probe of lower G/C content (e.g. some repetitive sequences) is used, or when you are working with heterologous probes.

3. The membrane is heat-sealed in a new plastic bag for subsequent detection.

6.2.2. Detection

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

1. Wash the membrane briefly with maleic acid buffer (buffer I) to remove any residues of SDS.

NOTE: To avoid unspecific binding of the antibodies, incubate the membrane for at least 60 minutes in maleic acid/1% (w/v) blocking reagent (buffer II), (1 ml/cm^2) on a shaker before adding antibody solution.

2. Dilute the antibody stock solution (750 U/ml) with buffer II to 75 mU/ml (1:10,000), (0.2 ml/cm^2).
3. Centrifuge the antibody stock solution before adding to the membrane in order to separate any precipitates, which can lead to background spots on the filter.
4. Discard buffer II and add the diluted antibody solution to the membrane.
5. Remove bubbles before sealing the bag.
6. Incubate for 30 minutes (no longer) at room temperature on a shaker.
7. Open the bag, remove buffer, and transfer membrane to a thoroughly cleaned dish with 5 ml/cm^2 wash buffer (buffer I plus 0.3% (v/v) Tween©20).
8. Wash 3 x 15 minutes with gentle agitation at room temperature.
9. Transfer membrane to a clean dish with alkaline buffer (buffer III) to activate the reporter enzyme alkaline phosphatase.

NOTE: The following detection methods are independent of the method utilized in the FAO/IAEA course, and might provide useful alternatives.

Colourimetric detection

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

The dye solution (10 ml/100 cm^2) is prepared by addition of 45 µl NBT-solution and 35 µl BCIP-solution to 10 ml buffer III. The incubation takes place in the dark for up to 20 hours. Avoid any shaking since this will cause a diffuse signal. The reaction can be stopped by washing the filter in TE buffer as soon the desired bands are visible.

Chemiluminescent detection

The chemiluminescence substrate AMPPD (or CSPD®) emits light after a two-step reaction. At first the molecule is de-phosphorylated by the enzyme alkaline phosphatase (AP) and in the second step the molecule decomposes and emits light. The emitted light appears as a continuous glow for more than 24 hours, and it can be documented on X-ray films. The advantages of chemiluminescence are remarkably improved sensitivity, the possibility to test different exposure times, and the facilitation of rehybridization experiments.

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

1. Dilute the CSPD® solution in buffer III to a final concentration of 0.235M (1:100), (1.5 ml/100 cm^2).
2. Place membrane on a clean transparent sheet and pipette the diluted CSPD® solution onto the membrane. Cover the membrane slowly with another transparent sheet to produce a uniform layer of liquid. Incubate for 5 minutes.
3. Place the membrane on 3 MM paper until the liquid is evaporated from the surface (do not let the membrane dry).
4. Seal the damp membrane in clingfilm and incubate for 15 minutes at 37°C.
5. Expose an X-ray film to the “glowing” membrane in the dark. The exposure times needed for genomic Southern blots are between 30 minutes and 14 hours.

6.3. Membrane rehybridisation method

1. For repeated hybridization of a membrane previously detected by chemiluminescence, wash it in sterile H_2O for 5 minutes.
2. Follow this by a 2 x 15 minutes incubation in 0.2 N NaOH, 0.1% SDS at 37°C in order to remove the bound Dig-labelled probe. After final washing in 2XSSC the filter is ready for new pre-hybridization.

6.4. References

Devos, K. M., M. D. Atkinson, C. N. Chinoy, H. A. Francis, R. L. Harcourt, R. M. D. Koebner, C. J. Liu, P. Masojc, D. X. Xie, and M. D. Gale**,** 1993. Chromosomal rearrangements in the rye genome relative to that of wheat. Theor.Appl.Genet. **85**: 673-680

Devos, K. M. and M. D. Gale**,** 2000. Genome Relationships: the grass model in current research. Plant Cell. **12**: 637-646

Feinberg, A. P. and B. Vogelstein**,** 1983. A technique for radiolabelling DNA restriction endonuclease fragments to a high specific activity. Anal.Biochem. **132**: 6-13

Hartl, D. L. and E. W. Jones. (1999) Essential Genetics. Jones and Bartlett Publishers, Sudbury, Massachusetts.

Kahl, G., 2001. The Dictionary of Gene Technology. Wiley-VCH, Weinheim.

Meinkoth, J. and G. Wahl**,** 1984. Hybridization of nucleic acids immobilized on solid supports. Anal.Biochem. **138**: 267-284

Sambrook, J., E. F. Fritsch, and T. Maniatis. (1989) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Southern, E. M.**,** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J.Mol.Biol. **98**: 503

6.5. Reagents needed

- Use only sterile-distilled water for all solutions
- 0.25M HCl. Concentrated HCl (37% (w/v) is 10 M, or 40x.
- NaCl
- Sodium citrate
- (5x) TBE per liter

TRIS base	54 g
Boric acid	27.5 g
EDTA 0.5 M	20 ml

- EDTA 0.2M
- LiCl 4M
- Ethidium bromide (EthBr)
- Antibody stock solution (750 U/ml) (Anti-Digoxigenin – alkaline phosphate) (provided in the Dig-Kit, Roche)
- Ammonium acetate 1M

- 70% Ethanol
- TE-buffer (10 mM Tris, 1mM EDTA, pH 8.0)
- CSPD solution in buffer III (alkaline buffer) (provided in the Dig-Kit, Roche)
- 0.2 N NaOH
- Maleic acid
- Tween®20
- Alkaline phosphatase (AP) (provided in the Dig-Kit, Roche)
- NBT/X-phosphate (provided in the Dig-Kit, Roche)
- AMPPD (resp. CSPD®) (provided in the Dig-Kit, Roche)
- Hexanucleotide-mixture (provided in the Dig-Kit, Roche)
- dNTP mixture (containing Dig-[11]-dUTP) (provided in the Dig-Kit, Roche)
- Klenow enzyme (DNA polymerase) (provided in the Dig-Kit, Roche)
- Bromophenol blue dye solution
  - 45 µl NBT-solution
  - 35 µl BCIP-solution
  - 10 ml buffer III
- NBT solution (75 mg/ml)*(BRL#95540) *Dissolved in dimethylformamide, TOXIC

BCIP solution
(50 mg/ml)*(BRL#95541) *Dissolved in dimethylformamide, TOXIC

- Denaturing solution
  - 0.4 M NaOH
  - 1M NaCl
- Loading buffer (x10) per ml

Glycerol (80%)	600 µl
Xylene cyanol	2.5 mg
Bromophenol blue	2.5 mg
H_2O	400 µl

- Hybridization pre-hybridization solutions (100 ml)

50% (v/v)	Formamide	(50 ml)
5% (w/v)	Blocking reagent	(5g)
5x	SSC (pH 7.0)	(25 ml 20xSSC)
0.1%	N-Lauroyl sarcosine	(1 ml of 10% stock)
0.02% (w/v)	SDS	(0.2 ml of 10% stock)

- Buffer I (Maleic acid buffer MAB)

0.1 M	Maleic acid (11.61 g/l)
0.15 M	NaCl (8.76 g/l)
pH 7.5	
Autoclave	

- Buffer II
  - Maleic acid/1% w/v)
  - Blocking reagent (provided in the Dig-Kit, Roche)

NOTE: It is advisable to prepare a 10 x concentrated stock solution of blocking reagent. Therefore, weigh 10 g of blocking reagent into an autoclavable flask, fill it up to ca. 90 ml

with buffer I, and heat it in an 80°C water bath to dissolve the blocking reagent (needs about 1 hour). The last particles can be dissolved by briefly boiling in a microwave. Autoclave the solution!

- Wash Buffer
  MAB + 0.3% (v/v) Tween©20
- Buffer III
  0.1 M TRIS-HCl (12.11 g/l)
  0.1. M NaCl (5.84 g/l)
  pH 9.5
  Autoclave

- 20% SDS
  Dissolve 200 g sodium dodecylsulphate in ddH_2O to final volume of 1 litre.
  You can use a low grade (Sigma #L5750) for hybridization washes, etc. and a better grade (Sigma #L4390) for hybridization solution, plasmid preps, stop solutions, *etc.*
- 20XSSC

NaCl	175.3 g
Na-citrate • 2 H_2O	88.2 g
Adjust pH to 7.4 with 1 N HCl	
Add H_2O to	1 litre

7. SSR

SSR (Microsatellite) definition: *Any one of a series of very short (2-10 bp), middle repetitive, tandemly arranged, highly variable (hypervariable) DNA sequences dispersed throughout fungal, plant, animal and human genomes (Kahl, 2001).*

Simple sequence repeats (SSR) or microsatellites are a class of repetitive DNA elements (Tautz and Rentz, 1984; Tautz, 1989). The di-, tri- or tetra-nucleotide repeats are arranged in tandem arrays consisting of 5 – 50 copies, such as $(AT)_{29}$, $(CAC)_{16}$ or $(GACA)_{32}$. SSRs are abundant in plants, occurring on average every 6-7 kb (Cardle et al., 2000). These repeat motifs are flanked by conserved nucleotide sequences from which forward and reverse primers can be designed to PCR-amplify the DNA section containing the SSR. SSR alleles, amplified products of variable length, can be separated by gel electrophoresis and visualised by silver-staining, autoradiography (if primers are radioactively labelled) or via automation (if primers are fluorescently labelled) (Figures 6.1 and 6.2). SSR analysis is amenable to automation and multiplexing (Figure 6.2), and allows genotyping to be performed on large numbers of lines, and multiple loci to be analysed simultaneously. SSRs can be identified by searching among DNA databases (e.g. EMBL and Genebank), or alternatively small insert (200-600bp) genomic DNA libraries can be produced and enriched for particular repeats (Powell et al., 1996). From the sequence data, primer pairs (of about 20 bp each) can be designed (software programmes are available for this).

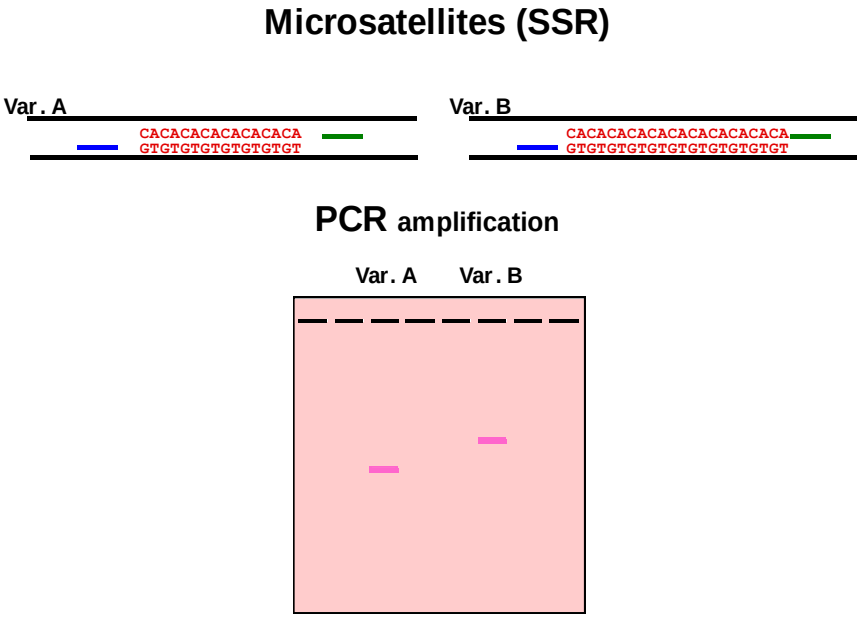


Figure 7-1**.** The schematic above shows how SSR variation (short A and long B) can be detected using gel electrophoresis after PCR with forward (blue) and reverse primers (green) (with permission, K. Devos).

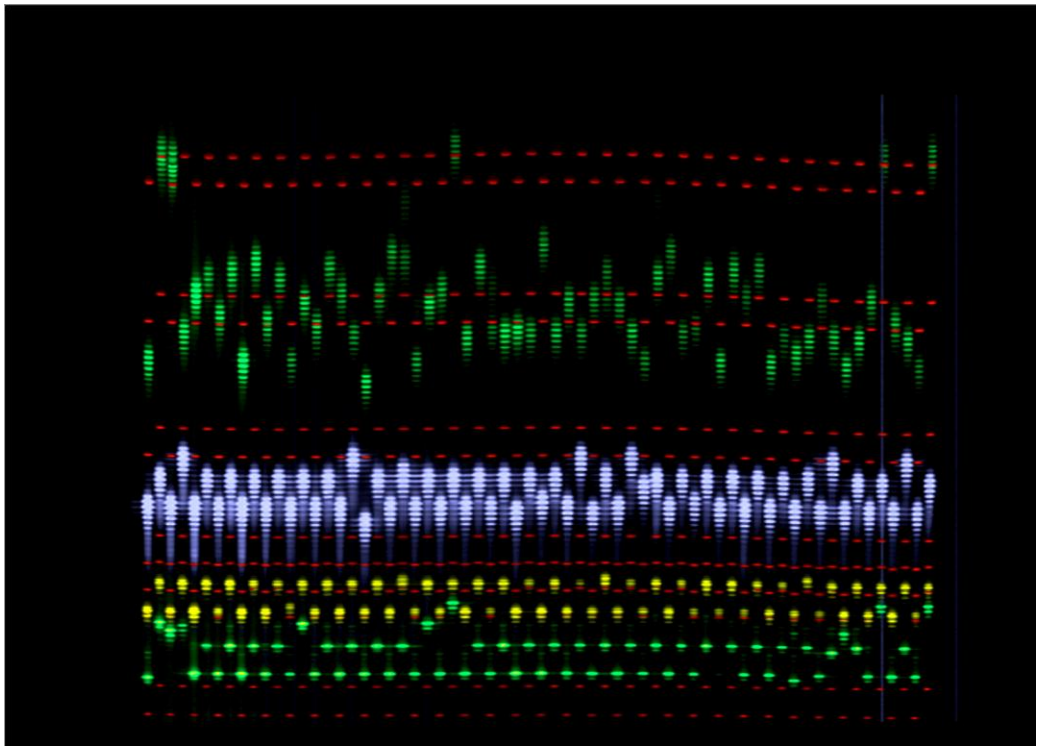


Figure 7-2. A computer image showing an example of SSR multiplexing with different colours (with permission, J. Kirby and P. Stephenson).

7.1. Protocol

7.1.1. PCR reaction mix

Microsatellite primers are specific for each individual genome or species. It is essential to know that the primer pairs chosen will work for your given species.

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

Prepare 25 µl Reaction Mix

1. Take four sterile PCR tubes and to each add the following:

10 x Taq buffer	2.5 µl
$MgCl_2$ (25mM)	1.5µl
dNTPs (10 mM)	1.0 µl
Forward primer (10 µM)	0.8 µl
Reverse primer (10 µM)	0.8 µl
Taq DNA polymerase (5U/µl)	0.25 µl
DNA (20ng/µl)	1.0 µl
*Add sterile distilled water up to	**25µl**

2. Mix by gently tapping against the tube.
3. Centrifuge briefly (~14,000 rpm for 5 seconds).

NOTE: Keep all reagents and reaction mix on ice until used.

7.1.2. PCR amplification

Place tubes in a PCR machine and amplify using a programme designed for the primers being used; an example is given below:

Step 1	Initial denaturing	94°C	5 minutes
Step 2	Denaturing	94°C	1 minute
Step 3	Annealing*	55°C	1 minute
Step 4	Extension	72°C	2 minute
Step 5	Cycling	repeat steps 2-5 for 34 cycles	
Step 8	Final extension	72°C	5 minutes
Step 9	Hold	4°C	forever

*NOTE: The annealing temperature (*Step 3*), in particular, can and does vary with primers used. Please note this when changing primers.

7.1.3. Separation of the amplification products in agarose gel

NOTE: Where SSR polymorphism is large, bands can be separated in agarose gels, however small base-pair differences among alleles require separation in polyacrylamide gels.

1. Take 5µl of the PCR product into a fresh tube.
2. Add 2 µl 5X loading buffer containing dye.
3. Centrifuge briefly (14,000 rpm for 5 seconds).
4. Load all 7µl of the mixture into a 1.5 % agarose gel (which is made up of 25% fine agarose and 75% normal agarose with 2µl/100ml ethidium bromide for staining DNA).
5. Run gel until dark blue colour marker has run two thirds of the gel.

NOTE: Do not run the dye off the gel or you will also lose your DNA samples.

NOTE: See Section of RFLP Protocol (Agarose gel electrophoresis) for details of gel preparation and running.

6. Stain gel with ethidium bromide (*Caution: ethidium bromide is toxic wear gloves and lab coat and avoid inhalation*).
7. Visualise under UV light (*Caution: wear gloves, and UV protective glasses or a shield over your face when you are exposed to the UV light of the transilluminator*).

7.1.4. Denaturing gel electrophoresis

NOTE: Denaturing the samples produces single-stranded DNA, which is used for detection in polyacrylamide gels (see below). Single-stranded detection is preferred as it results in a greater clarity in band separation for detection. Setting up and casting a polyacrylamide gel using sequencing apparatus involves the followings.

7.1.5. Assembling the glass plate sandwich

1. Wear gloves and lab coat, and place the Integral Plate Chamber (IPC), *i.e.* the big plate on the bench, horizontally, glass side up. Clean the upper surface of the glass plate using Alconox and warm water. Rinse and dry the plate.
2. Clean the upper surface with 95% ethanol. Apply a thin film of Sigmacote (2ml) to the upper surface of the plate and spread evenly using blue roll and dry. Repel silane or Repelcote are other brand names of the same product.

NOTE: Change gloves between working with the bigger and smaller plates as you will be using 2 different chemicals, bind silane and repel silane that must not contaminate the unintended glass plate. One is a 'binder' while the other repels and when properly applied ensure that the gel sticks only one surface and not the other. Contamination can be brought about by not changing gloves and this will lead to breakage of the gel between the 2 plates!

3. Clean the smaller plate using Alconox and water (you may also need to use a razor blade to remove old bits of gel that have stuck). Rinse and dry the plate, clean the upper surface only with 95% ethanol.
4. Prepare fresh bind silane solution by adding 3µl of binding solution to 1ml of 95% ethanol mixed with 5µl of glacial acetic acid.
5. Apply prepared bind silane solution to the upper surface of the plate and spread evenly using blue roll.

NOTE: Clean everything following use, and dispose of materials carefully according to the regulations of your organization.

NOTE: The glass plates must be meticulously clean. Detergent microfilm left on the glass plate may result in a high (brown coloured) background for the stained gel.

6. Place clean, dry spacer on the long edge of the IPC plate. Make sure that there is no untrimmed adhesive underneath the spacer.
7. Place the outer glass plate on the top of the spacers. The raised plastic edges on the IPC plates will help position the spacer and plate. Align the outer plate and spacer with bottom edge. Precise alignment is necessary.
8. Slide clamps over the gel plate assembly, one clamp at a time. This can be done while holding the IPC vertically. Start each clamp (there is right and left clamp) near the bottom

end first, then slide the clamp on to the IPC assembly until it snaps into a place along the entire length.

NOTE: The clamps must fit reasonably tightly to prevent the spacer from leaking. Make sure the clamps are all the way on, with the spacer and outer glass plate flush at the bottom.

7.1.6. Casting gel

1. Prepare 100 ml of gel solution per plate by adding together:

*Acrylamide/bis solution 19:1 (40 %)	15 ml
TBE (10X)	10 ml
Urea 8 M	50 g
Make up to 100 ml with distilled water.	

* *Caution: acrylamide is toxic*

NOTE: An alternative option is to use a pre-mixed solution, SequaGel®XR (National Diagnostics, Inc.), which gives sharper bands

2. Filter the solution and keep at 4°C and take as required when ready to cast a gel.
3. Add 28 µl TEMED (*Caution: TEMED is corrosive*) and 800 µl 10% fresh ammonium persulphate solution (*Caution: ammonium persulphate, APS, is harmful*) to 100 ml of the gel mix,
4. Gently draw up acrylamide solution into a 100ml syringe, avoiding air bubbles.
5. Adjust angle of plates so gel solution flows slowly down one side. Keep the acrylamide solution flow consistent by varying the flow rate by tilting the gel assembly. This reduces the formation of bubbles during the filling. Perfect clean plates will not allow bubbles to form. If bubbles do form, tap the glass plate gently to dislodge them.

NOTE: Gel will start to polymerize after adding APS, be prepared to move quickly.

6. Insert the flat side of a 0.4mm shark's tooth comb between plates before the gel polymerizes. Place the binder clamps over the glass plates to insure that the plates are held firmly against the comb
7. Leave to polymerise for approximately 1 hour.

NOTE: Make up the developer for silver-staining while the gel is polymerising, see section 5.1.6 below.

7.2. Setting up the operation

1. Place the IPC assembly into the universal base, against the back of the wall. Stick a gel temperature indicator on to the outer plate, somewhere near the centre of the gel, to monitor the temperature during electrophoresis.
2. Fill the upper buffer chamber with 1X TBE buffer. The level of the buffer should be about 1cm from the top all the time during the run.

3. Fill the lower buffer chamber and adjust the levelling screws. Do not fill the lower chamber with more than 500ml of buffer
4. Remove the comb from the gel and clean the well space using distilled water. Replace comb carefully, teeth first this time.

NOTE: You can only replace the comb once, so be very careful!

5. Pull the plastic hood over the gel tank and insert the electrodes. Switch on the power pack and adjust the reading roughly to 900-1500 V and 70W.
6. Pre-run the gel at 125 watts. The gel temperature will stabilize near 55°C. Pre-running the gel at 45°C for an hour or two may result in better resolution, particularly if you use high catalyst concentration

7.3. Polyacrylamide gel running conditions

1. Prepare samples by adding 2 µl of formamide dye mix to 8 µl of your PCR reaction (second half). Denature the samples for 5 minutes and place on ice (Caution: formamide is harmful).
2. Load 1 kb marker ladder (10 µl 1 kb ladder (50 ng/µl) add 6 µl formamide loading buffer); load 5 µl into first lane (and at convenient intervals across the gel).
3. Load 8 µl of each sample containing the formamide dye mix into individual wells of the gel.
4. Run gel for approximately 1 hour and 20 minutes at 75 watts or until just before the dark blue runs off the bottom of the gel. You will need to quantify the best time for your particular PCR products.

NOTE: Do not run the dye off the gel or you will also run your sample off the gel and lose it.

7.4. Silver-staining

1. While the gel is polymerising, prepare the developer solution: Dissolve 60 g sodium carbonate in 2 litres of distilled water then add 400 µl of sodium thiosulphate solution (10 mg/ml) and 3 ml formaldehyde (37% solution) and store at 4°C (*Caution: Both sodium carbonate and formaldehyde are toxic, avoid inhalation and wear gloves and lab coat).For best results, the developer must be chilled.*
2. While the gel is running, prepare the fixative (10 % acetic acid): Add 200 ml glacial acetic acid to 1.8 litres distilled water (*Caution: acetic acid is corrosive, gloves should be worn*).
3. Prepare the silver-stain (*toxic, wear gloves*): Add 2g silver nitrate ($AgNO_3$) solution in 2 litres of distilled water (*Caution: silver nitrate is corrosive, gloves should be worn*). Then add 3 ml formaldehyde (37% solution) and mix (*Caution: formaldehyde solution is toxic, Wear gloves and lab coat, and avoid inhalation*). Silver nitrate is light sensitive so store in an opaque bottle or wrap aluminium foil around the bottle.
4. Remove the gel from the rig and separate the plates. Place the gel in a tray with the fixative and leave shaking in a fume hood for 20 minutes.

NOTE: Do not pour solutions directly onto the gel as it may come off the plate! When running

5. Remove the gel and stand on a rack. Pour off fixative and save it as it can be used for up to 10 times. Wash the gel three times (2 min) in water. Remove the gel and stand. Pour out the water and replace with silver-stain, introduce the gel again and leave shaking for 30 minutes. For best results, cover the tray as light affects the $AgNO_3$ solution

NOTE: Silver stain ($AgNO_3$ and formaldehyde solution) can be re-used up to 10 times

NOTE: The next few procedures have to be followed quickly and carefully so make sure you have everything set up and ready.

6. Remove gel from the silver-stain solution and rest it on a tray containing water (do not put it in the water yet). Dispose of spent stain according to the regulations of your organization. Rinse the box that contained the silver-stain with water.
7. Set a timer for 10 seconds. Start the timer and quickly lower the gel into the water. Agitate several times to remove all excess silver-stain. When 10 seconds is up quickly drain the gel and place it in the developing solution.
8. Agitate the gel in developer solution and, use a piece of white paper placed behind the gel to check progress of the band development. Keep an eye on the gel as it develops. Stop the reaction when bands start to appear near the bottom of the gel (*i.e.*: 70 bp marker on the 1 kb ladder) by taking the gel out of the developer solution.
9. Put the gel into a tray containing 2 litres of stop solution (10% glacial acetic acid) for 5 minutes.

NOTE: The stop solution could be what was saved from earlier (first step fixative) if there is no need for re-use. If re-use is desired, it is best to have separate fixative and stop solutions as the latter contains AgNO3 and therefore not suitable for use again as fixative.

10. Rinse gel in water for 5 minutes and leave it to dry standing vertically.
11. Gels can be recorded or documented using Kodak duplicating film.
    11.1. Place the glass plate upside down on the film.
    11.2. Expose to room light for 15-17 seconds (depending on the room light intensity).

NOTE: The longer the light exposure, the brighter the film gets following development.

Gels can be scanned or photocopied.

7.5. References

Cardle, L., L Ramsay, D. Milbourne, M. Macaulay, D. Marshall, and R. Waugh**,** 2000. Computational and experimental characterisation of physically clustered simple sequence repeats in plants. Genetics. **156**: 847-854.

Kahl, G., 2001. The Dictionary of Gene Technology. Wiley-VCH, Weinheim.

Powell, W., G. C. Machray, and J. Provan**,** 1996. Polymorphism revealed by simple sequence repeats. Trends in Plant Sci. **1**(7): 215-222.

Tautz, D.**,** 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Res. **17**: 6463-6471

Tautz, D. and M. Rentz**,** 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nature. **322**: 652-656.

7.6. Reagents needed

- Use only sterile distilled water for all solutions.
- *Taq* buffer
- dNTPs
- Alconox
- Repel silane (Repelcote, Sigmacote)
- Bind silane
- Sterile distilled water
- Primers
- *Taq* DNA polymerase (5U/µl)
- DNA (10-20ng/µl)
- 10 x loading buffer

Glycerol (80%)	600 µl
Xylene cyanol	2.5 mg
Bromophenol blue	2.5 mg
Distilled water	400 µl

- 5 x loading buffer

Glycerol (80%)	300 µl
Xylene cyanol	1.3 mg
Bromophenol blue	1.3 mg
Distilled water	400 µl

- Ethidium bromide
- Agarose
- Acrylamide
- Bis-acrylamide
- TEMED
- Ammonium persulphate
- Sodium thiosulphate
- TBE

H_2O	~800 ml
Tris base	108 g
Boric acid	55 g
EDTA	9.3 g
ddH_2O	Adjust volume to 1 litre

- 100% ethanol
- Bind silane

- Sodium carbonate
- Glacial acetic acid

- Formamide dye mix (for 1 ml)

Formamide (deionized)	950µl
dd H_2O	30µl
EDTA (0.5 M)	20µl
Bromophenol blue	1 mg
Xylene cyanol	1 mg

Mix and store at -20°C

8. ISSR

ISSR amplification definition: A variant of the polymerase chain reaction that uses simple sequence repeat primers (e.g. $[AC]_n$) to amplify regions between their target sequences (Kahl, 2001).

Inter-SSR (ISSR) amplification is an example (one of many) of a PCR-based fingerprinting technique. The technique exploits the abundant and random distribution of SSRs in plant genomes by amplifying DNA sequences between closely linked SSRs (Figure 6.1). The method used in the FAO/IAEA course used 3'-anchored primers to amplify regions between two SSRs with compatible priming sites (Yang et al., 1996). More complex banding patterns can be achieved using 5'-anchored primers that incorporate the SSR regions in their amplification products, and by combining 3'- and 5'- primers (Zietkiewicz et al., 1994).

Other methods of fingerprinting using primers complementary to SSR motifs involve using SSR specific primers in combination with an arbitrary primer (Davila *et al.*, 1999)~~(Davila *et al.*, 1999)~~, or in combination with primers that target other abundant DNA sequences such as retrotransposons (Provan *et al.*, 1999)~~(Provan *et al.*, 1999)~~.

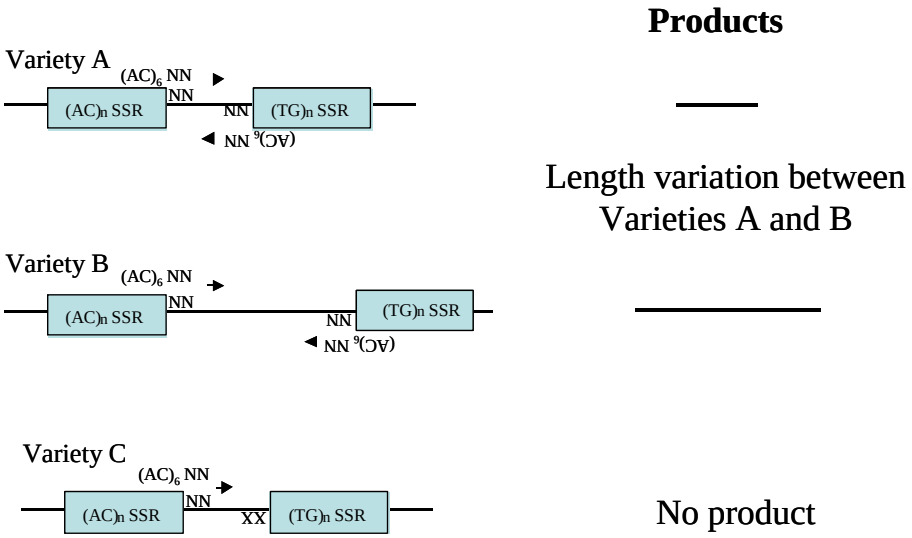


Figure 8-1. The above scheme shows how sequence variation between two SSRs results in variation in PCR products in varieties A, B and C. The figure shows variation at only one ISSR locus, amplification of all compatible ISSR loci among the genomes of a range of varieties will result in complex, fingerprinting, banding patterns.

8.1. Protocol

In the example below, one of three primers given in the ISSR protocol of Yang *et al.*, (1996) is used; this produces a relatively simple fingerprint (small number of bands). In more recent applications two or more primers have been used to produce more complex banding profiles (similar to AFLP profiles).

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

8.1.1. Prepare 20µl reaction mix

1. Take one PCR tube and add:

10x PCR buffer	2.5 µl
$MgCl_2$ (25mM)	1.5µl
Primer (10 mM)	2.5 µl
dNTPs (10mM)	0.8 µl
DNA (20ng/ µl)	1.25 µl
Taq DNA polymerase (5 U/ µl)	0.2 µl
Add sterile distilled water to bring volume to 20 µl	

2. Mix by tapping bottom of tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds)

NOTE: Keep all reagents and reaction mix on ice.

8.1.2. PCR amplification

Place tube in a PCR machine and amplify using a programme designed for the primer(s). In this example the following programme can be used:

Step 1	Initial denaturing	94°C	7 minutes
Step 2	Denaturing	94°C	30 seconds
Step 3	Annealing*	54°C	45 seconds
Step 4	Extension	72°C	2 minute
Step 5	Cycling	repeat steps 2-4 for 30 cycles	
Step 6	Final extension	72°C	7 minutes
Step 7	Hold	4°C	forever

8.1.3. Separation and visualization of the amplification products

1. Add 2 µl of 5x loading buffer to 8 µl of PCR sample.
2. Vortex briefly.
3. Centrifuge briefly (14,000 rpm for 5 seconds)
4. Load samples into a non-denaturing 6% polyacrylamide gel/3M urea gel (see Section 5.1.4. of SSR protocol for preparation of 6% acrylamide gel. [Step 4: Use 180 g urea (3M) instead of 480 g (8M)!])

NOTE: Where the running of polyacrylamide gels is not feasible, 1.5% agarose gel may be used for fragment separation. For this, load sample into 1.5% agarose gel. A mixture of 25% fine

agarose and 75% routine agarose works very well (see Section 6.1.3. of SSR protocol for preparation of agarose gel [Step 4]).

8.1.4. Gel running conditions

1. Run gel under non-denaturing condition at 12 V/cm for 10-13 hours.

NOTE: This is normally done overnight.

NOTE: Non-denaturing gels are run at low voltages and 1 x TBE to prevent denaturation of small fragments of DNA by the heat generated in the gel during electrophoresis.

2. Run agarose gel at 120V for at least 2 hours

NOTE: Do not run the bands off of the bottom of the gel.

8.1.5. Silver-staining

Follow Section 6.1.6 of SRR Protocol (silver-staining).

8.2. Primers available at Plant Breeding & Genetics Laboratory (FAO/IAEA)

Pimers ID	Sequence information	Pimers ID	Sequence information
ISSR-1	$(CAC)_7$ T	ISSR-27	$(GT)_8$ G
ISSR-2	$(GA)_9$C	ISSR-28	$(AC)_8$ T
ISSR-3	GT)$_9$G	ISSR-29	$(AC)_8$ C
ISSR-4	$(CAC)_7$G	ISSR-30	$(AC)_8$ G
ISSR-5	GT$(CAC)_7$	ISSR-31	$(TG)_8$ A
ISSR-6	GTG)$_7$C	ISSR-32	$(TG)_8$ G
ISSR-7	$(CA)_{10}$G	ISSR-33	AG)$_8$ YT
ISSR-8	$(CT)_9$G	ISSR-34	$(GA)_8$ YT
ISSR-9	$(GA)_9$AY	ISSR-35	$(CT)_8$ RA
ISSR-10	BDB$(TCC)_5$	ISSR-36	$(CT)_8$ RC
ISSR-11	HVH$(TCC)_5$	ISSR-37	$(CA)_8$ RT
ISSR-12	$(AG)_8$ T	ISSR-38	$(CA)_8$ RC
ISSR-13	$(AG)_8$ G	ISSR-39	$(GT)_8$ YA
ISSR-14	$(GA)_8$ T	ISSR-40	$(GT)_8$ YG
ISSR-15	$(GA)_8$ C	ISSR-41	$(TC)_8$ RT
ISSR-16	$(GA)_8$ A	ISSR-42	$(AC)_8$ YG
ISSR-17	$(CT)_8$ A	ISSR-43	$(AC)_8$ YA
ISSR-18	$(CT)_8$ G	ISSR-44	$(AC)_8$ YT
ISSR-19	$(CT)_8$ T	ISSR-45	$(TG)_8$ RT

ISSR-20	$(CA)_8$ A	ISSR-46	$(TG)_8$ RC
ISSR-21	$(CA)_8$ G	ISSR-47	$(ACC)_6$
ISSR-22	$(GT)_8$ A	ISSR-48	$(ATG)_8$
ISSR-23	$(GT)_8$ C	ISSR-49	$(CTC)_6$
ISSR-24	$(GT)_8$ T	ISSR-50	$(GAA)_6$
ISSR-25	$(TC)_8$ A	ISSR-51	$(GACA)_6$
ISSR-26	$(GT)_8$ C	ISSR-52	$(TCC)_5$ RY

Y=C/T
R=A/G

8.3. References

Davila, J. A., Y. Loarce, and E. Ferrer, 1999. Molecular characterization and genetic mapping of random amplified microsatellite polymorphism in barley. Theor.Appl.Genet. 98: 265-273

Provan, J., W. T. B. Thomas, B. P. Forster, and W. Powell, 1999. Copia-SSR: a simple marker technique which can be used on total genomic DNA. Genome. 42: 363-366

Yang, W., A. C. De Olivera, I. Godwin, K Schertz, and J. L. Bennetzen, 1996. Comparison of DNA marker technologies in characterizing plant genome diversity: variability in Chinese sorghums. Crop Sci. 36: 1669-1676

Zietkiewicz, E., A. Rafalski, and D. Labuda, 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored Polymerase Chain Reaction Amplification. Genomics. 20: 176-183

8.4. Reagents needed

Use only sterile distilled water for all solutions:

- *Taq* buffer
- dNTPs
- Sterile distilled water
- Primer(s)
- *Taq* DNA polymerase (5U/ µl)
- DNA (10-20 ng/ µl)
- 10 x loading buffer
  - Glycerol (80%) 600 µl
  - Xylene cyanol 2.5 mg
  - Bromophenol blue 2.5 mg
  - Water 400 µl
- 5 x loading buffer
  - Glycerol (80%) 300 µl
  - Xylene cyanol 2.5 mg
  - Bromophenol blue 2.5 mg
  - Water 400 µl
- Ethidium bromide

- Agarose
- Acrylamide
- Bis-acrylamide
- TEMED
- Ammonium Persulphate
- Alconox
- TBE (see 5.3)
- Ethanol(95%)
- Repelcote (Symacote)
- Bind silane
- Sodium carbonate
- Glacial acetic acid
- Sodium thiosulphate

9. AFLP

Amplified Fragment Length Polymorphism (AFLP) is basically a fingerprinting technique. It is a method by which selection of restricted fragments of a total genomic DNA digest is detected by PCR amplification. It is a combination of hybridisation and amplification-based strategies.

The AFLP technique combines components of RFLP analysis with PCR technology (Vos *et al.*, 1995)~~(Vos *et al.*, 1995)~~. Total genomic DNA is digested with a pair of restriction enzymes, normally a frequent and a rare cutter. Adaptors of known sequence are then ligated to the DNA fragments. Primers complementary to the adaptors are used to amplify the restriction fragments. The PCR-amplified fragments can then be separated by gel electrophoresis and banding patterns visualized (Figure 7.1). A range of enzymes and primers are available to manipulate the complexity of AFLP fingerprints to suit application. Care is needed in selection of primers with selective bases.

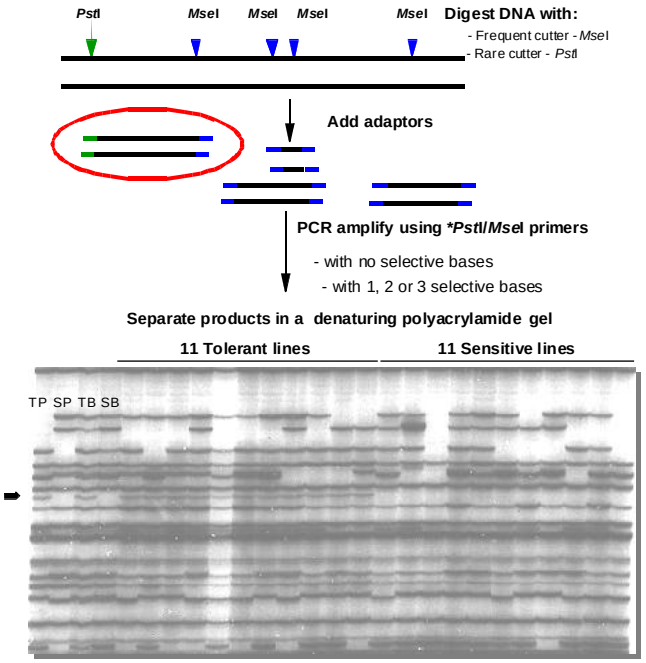


Figure 9-1**.** In the figure above AFLP profiles have been used in bulk segregant analysis to detect a band associated with tolerance to aluminium in rye, the arrow shows the presence or absence of a band in the tolerant (TP) and susceptible (SP) parents, tolerant (TB) and susceptible (SB) bulks, and 11 tolerant and 11 susceptible individuals (scheme and data with permission, K. Devos and Miftahudin, respectively).

9.1. Protocol

AFLP involves four major steps:
I* Cutting genomic DNA with restriction enzymes
II* Ligating double-strand adaptors to the restriction fragments
III Amplifying (pre- and selective amplification) restriction fragments using primers
IV Gel analysis of the amplified products

*OPTIONAL: these two steps can be performed in one reaction

NOTE: Wear gloves and lab coat at all times for safety and to prevent contamination.

9.1.1. Restriction of genomic DNA and ligation of adapters to the DNA fragments

Two pairs of restriction enzymes, *Mse*I/*Tru*91and *Pst*I/*Eco*RI, were used to digest the genomic DNA. *Mse*1/*Tru*91 is a frequent cutter with a T↓TAA cutting site, whereas *Pst*I and *Eco*RI are 6-base rare cutters with a CTGCA↓G (*Pst*I is methylation sensitive) and G↓AATTC (*Eco*RI)

1. Put on gloves (to protect yourself and the reaction mix) and add the following to a 0.5 ml Eppendorf tube:

Restriction-ligation reaction mixture

Genomic DNA(20ng/µl)	~150ng
5x RL buffer	2µl
Rare cutting enzyme EcoRI (10U/µl)	0.10 µl
Frequent cutting enzyme Tru91 (10U/µl)	0.10 µl
EcoRI adaptor mix (50 pmole/µl)	0.5 µl
Tru9I adapter mix (50 pmole/µl)	0.5 µl
rATP (10 mM)	0.2 µl
T4 DNA ligase (5U/µl)	0.13µl
Sterile distilled water	Up to 10µl

2. Mix by tapping the bottom of the tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds).
4. Incubate the resulting reaction mixture for a minimum of 3 hours at 37°C.
5. Inactivate the restriction endonuclease by incubating the mixture at 70°C for 15 min.
6. Place tubes on ice and do brief centrifugation to collect contents.

9.1.2. Pre-amplification

Pre-amplification is performed with primers having one selective nucleotide. The aim of pre-amplification is to generate enough template DNA for selective amplification step.

1. Set up the PCR reaction (on ice)

10 x PCR buffer	5 µl
Restriction-ligation reaction (from 7.1.1)	5 µl
EcoRI primer (10µM/µl)	1.5 µl
Mse1/Tru91 primer (10µM/µl)	1.5 µl
dNTPs (10 mM)	1 µl
Taq DNA polymerase (5U/µl)	0.5 µl
Sterile distilled water	Up to 50µl

2. Mix by tapping the bottom of the tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds).

NOTE: The *Eco*RI and *Tru*91 primers used in pre-amplification are non- selective in that they recognise all *Eco*RI and *Tru*91 priming sites.

9.1.3. PCR pre-amplification

This step amplifies all of the DNA fragments carrying PstI and TruI terminal adaptors, and provides sufficient template for subsequent selective amplification.
Place the tube in the PCR machine and amplify using the following programme:

Step 1	Denaturing	94°C	30 seconds
Step 2	Annealing	65°C (-0.7 °C/cycle)	30 seconds
Step 3	Extension	72°C	1 minute
Step 4	Cycling	repeat steps 1-3 for 11 cycles	
Step 5	Denaturing	94°C	30 seconds
Step 6	Annealing	56°C	30 seconds
Step 7	Extension	72°C	1 minute
Step 8	Cycling	repeat steps 5-7 for 22 cycles	
Step 9	Hold	4°C	forever

9.1.4. Check-step

It is important to check that everything has worked in the previous steps before proceeding.

1. Take a 5 µl aliquot of the PCR-amplified product from 7.1.3 above and place in a fresh 0.5 ml tube, and add 2 µl 5x loading buffer.
2. Vortex briefly.
3. Centrifuge briefly (14,000 rpm for 5 seconds).
4. Load the sample into a 1.2 % agarose gel.
5. Run gel at 50V for 30 minutes.

6. Visualise DNA by UV illumination (Figure 8.2).

(*Caution: wear gloves, and UV protective glasses and shields over your face when you are exposed to the UV light of the transilluminator*)

NOTE: If previous steps have worked you should see a clear DNA band (Figure 8.2).

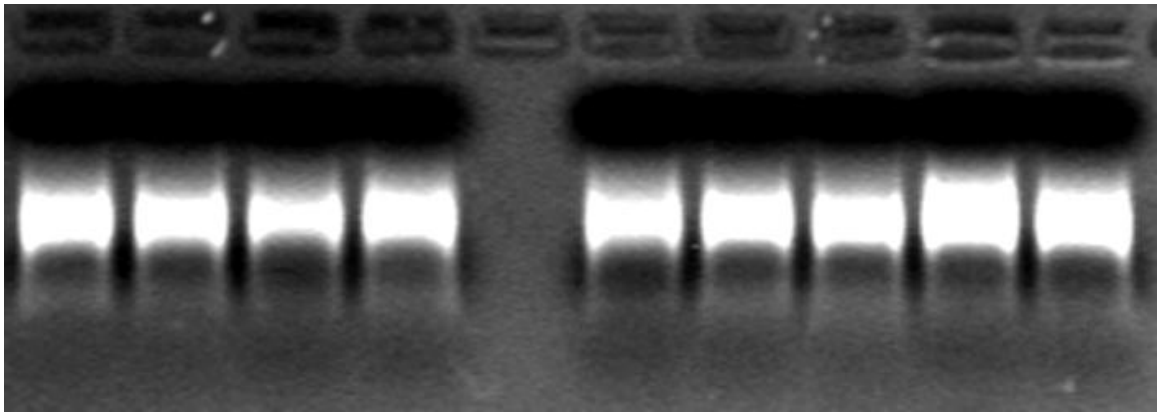


Figure 9-2.

7. Dilution of pre-amplified DNA:
   - For silver staining, dilute 5µl of pre-amplified DNA sample 1:50 with water (50 µl sample + 245 µl water).
   - For fluorescent labelling, dilute pre-amplified DNA to 1:10 with TE (10 µl sample + 90 µl water).
   - Store this dilution and the remaining pre-amplification product at -20°C (long term).

   NOTE: The dilution of sample depend of amplified products (S.7.) that is used in selective amplification (8.1.3) PCRs, and now termed 'Test DNA'.

9.1.5. Selective pre-amplification

In this section, specific subsets in the test DNA are amplified using EcoRI and Tru91 primers that are extended with one to three selective nucleotides. Silver staining of the amplified fragments that have been electrophoresed on PAGE is commonly used for detection of DNA banding patterns. Alternatively, fluorescence-labelled primers can be used in the selective amplification PCR step and the products visualised on an automated DNA analyser. These two options are described below.

9.1.6. PCR mix for selective amplification, products to be visualized on PAGE

1. Put on gloves and in a PCR tube add:

Test DNA (diluted pre-amplified DNA from 8.1.4:Step 7)	5.0 µl
10 x PCR buffer	2.5 µl
*Eco*RI selective primer (10 µmol)	0.25 µl
*Tru*91 selective primer (10 µmol)	0.75 µl
dNTPs (10 mM)	0.5 µl
Taq DNA polymerase (5U/µl)	0.2 µl
Sterile distilled water	Up to **25.0µl**

2. Mix by gently tapping against the tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds).

9.1.7. PCR profile for Selective amplification, products to be visualised on PAGE.

Place tube in the PCR machine and amplify using the following programme:

Step 1	Denaturing	94°C	30 seconds
Step 2	Annealing	65°C (-0.7 °C/cycle)	30 seconds
Step 3	Extension	72°C	1 minute
Step 4	Cycling	repeat steps 1-3 for 13 cycles	
Step 5	Denaturing	94°C	30 seconds
Step 6	Annealing	56°C	30 seconds
Step 7	Extension	72°C	1 minute
Step 8	Cycling	repeat steps 5-7 for 23 cycles	
Step 9	Hold	4°C	forever

9.1.8. Polyacrylamide Gel Electrophoresis (PAGE)

The single-stranded AFLPs are separated in long, denaturing polyacrylamide gels (often referred to as sequencing gels).

1. Take a 5 µl aliquot of the PCR-amplified product from 10.1.3 above and place in a fresh 0.5 ml tube, and add 2 µl formamide loading buffer. The number of samples will be determined by the number of wells you have in your polyacrylamide gel.
2. Denature for 5 minutes at 95°C - 100°C, and snap-cool on ice.
3. Centrifuge briefly (14,000 rpm for 5 seconds).
4. Run 5µl samples in denaturing 6% polyacrylamide gels. SequaGel®XR (http://www.nationaldiagnostics.com/electroproducts/ec842.html)

9.1.9. Silver staining of PAG

Follow the procedure given in the SSR Protocol (6.1.6. Silver-staining).

9.1.10. PCR mix for selective amplification, products to be visualized on an automated DNA analyser

1. Put on gloves and in a PCR tube add:

Test DNA (diluted DNA from .7.1.2.2:S7)	5.0 µl
10 x PCR buffer (with Mg2+)	2.0 µl
Fluorescent EcoRI Primer (1µmol)	1.0µl
*Tru*91 selective primer (5µmol)	1.0µl
dNTPs (10 mM)	0.40µl
Taq DNA polymerase (5U/µl)	0.20 µl
Sterile distilled water up to	**20.0µl**

2. Mix by gently tapping against the tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds).

9.1.11. PCR profile for selective amplification, products to be visualized on an automated DNA analyser

Place tube in the PCR machine and amplify using the following programme:

Step 1	Denaturing	94°C	30 seconds
Step 2	Annealing	65°C (-0.7 °C/cycle)	30 seconds
Step 3	Extension	72°C	1 minute
Step 4	Cycling	repeat steps 1-3 for 11 cycles	
Step 5	Denaturing	94°C	30 seconds
Step 6	Annealing	56°C	30 seconds
Step 7	Extension	72°C	1 minute
Step 8	Cycling	repeat steps 5-7 for 29 cycles	
Step 9	Hold	4°C	forever

9.1.12. Electrophoresis using an automated DNA analyser

The single-stranded AFLPs are separated through electrophoresis on a capillary type automated DNA analyser (ABI Prism 3100 is used in the Plant Breeding and Genetics Laboratory).

1. Put on gloves and in a "sequencer" plate, add for each sample:

PCR-amplified product from 7.1.3.1	1.0µl
Formamide	13.0 µl
ROX standard	0.25 µl

2. Denature for 5 minutes at 95°C - 100°C, and snap-cool on ice.
3. Centrifuge briefly (14,000 rpm for 5 seconds) and check for air bubbles.
4. Load plate on the DNA analyser according to User's manual and select the option for AFLP fragment separation.

9.1.13. Production of single primer, linear PCR products

NOTE: This procedure is used to avoid doubled stranded DNA fragments and results in a greater clarity of band separation.

1. Put on gloves and add in a PCR tube:

10X PCR buffer	2 µl
Selective amplification DNA (produced in Step 6)	2 µl
*Pst*I selective primer (50 ng/µl)	1.5 µl
dNTPs (2 mM)	2.5 µl
Taq DNA polymerase (5U/µl)	0.1 µl
Add sterile distilled water to make up to	**20 µl**

2. Mix gently by tapping the tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds).

9.1.14. PCR amplification to produce single stranded DNA

Put on gloves and place tube from 10.2..3. into a PCR machine and amplify using the following programme:

Step 1	Denaturing	94°C	30 seconds
Step 2	Annealing	56°C	30 seconds
Step 3	Extension	72°C	1 minute
Step 4	Cycling	repeat steps 1-3 for 22 cycles	
Step 5	Denaturing	94°C	30 seconds
Step 6	Hold	4°C	hold

9.2. Required enzymes and primer sequences for AFLP assays

9.2.1. Restriction enzymes

MseI/Tr91

```
5’  T | T  A  A  3’
3’  A  A  T | T  5’
```

PstI

```
5’  C  T  G  C  A | G  3’
3’  G | A  C  G  T  C  5’
```

9.3. Preparation of adapters

Tru9I adapter-oligos have 16 and 14 nucleotides

5’- GACGATGAGTCCTGAG-3’
3’-TACTCAGGACTCAT-5’

Take 15µl of each to get the final concentration of 50pmol/µl in 30µl water.

Pst1 adapter-oligos have 21 and 14 nucleotides

5’- CTCGTAGACTGCGTACATGCA -3’
3’-CATCTGACGCATGT-5’

Take 15µl of each to get the final concentration of 50pmol/µl in 30µl water.

9.4. Reagents needed

- Use only sterile distilled water for all solutions
- 5x RL buffer
  - 50 mM TrisAc pH7.5
  - 50 mM MgAc
  - 250 mM KAc
  - 25 mM DTT
  - 250 ng/µl BSA
- Rare cutting enzyme, *Pst*I (5U/µl)
- Frequent cutting enzyme, *Tru91* (5U/µl)
- *Pst*I adaptor (5 pmole/µl) or *Eco*RI adaptor (5 pmole/µl)
- *Tru9I* adaptor (50 pmole/µl)
- rATP (10 mM)
- T4 DNA ligase
- 10 x PCR buffer
- *Pst*I or *Eco*RI non-selective primer (50 ng/µl)
- *Tru9I* non-selective primer (50 ng/µl)
- *Taq* DNA polymerase (5U/µl)

- Agarose
- $T_{0.1}E$ buffer
- *Pst*I or *Eco*RI selective primer
- ***Tru9I*** selective primer
- dNTPs (10 mM)
- Formamide
- ROX Standard

9.5. Sequence information of adapters and primers used for AFLP

*Tru*91-Adapter sequence l	5'-GACGATGAGTCCTGAG-3'
*Tru*91-Adapter sequence l.	3'-TACTCAGGACTCAT-5'
*Eco*RI:	5'- CTCGTAGACTGCGTACC -3'
*Eco*RI	5'- AATTGGTACGCAGTCTAC -3'
Primers for pre-amplification	
*Tru*91-primer	5'-GACGATGAGTCCTGAGTAA-3'
Eco-P0:	5'- GACTGCGTACCAATTC -3'
*Tru*91-P0:	5'- GATGAGTCCTGAGTAA -3'
*Tru*91-PC:	5'- GATGAGTCCTGAGTAAC -3'
*Tru*91 Selective primers**	
*Tru*91-CAC	5-GATGAGTCCTGAGTAACAC-3'
*Tru*91-ACC	5'-GATGAGTCCTGAGTAAACC-3'
*Tru*91-CCA	5'-GATGAGTCCTGAGTAACCA-3'
*Tru*91-CAA	5'-GATGAGTCCTGAGTAACAA-3'
*Tru*91-ACG	5-GATGAGTCCTGAGTAAACG-3'
*Tru*91-CAG	5'-GATGAGTCCTGAGTAACAG-3'
*Tru*91-CAT	5'-GATGAGTCCTGAGTAACAT-3'
Tru91-CGA	5'-GATGAGTCCTGAGTAACGA-3'
*Tru*91-CGT	5'-GATGAGTCCTGAGTAACGT-3'
*Tru*91-CCT	5'-GATGAGTCCTGAGTAACCT-3'
*Tru*9I -CTA-	5'- GATGAGTCCTGAGTAACTA 3'
*Tru*9I -CTC	5'- GATGAGTCCTGAGTAACTC -3'
*Tru*9I -CTG:	5'- GATGAGTCCTGAGTAACTG -3'
*Tru*9I -CTT:	5'- GATGAGTCCTGAGTAACTT -3'
*Tru*9I -GAA	5'- GATGAGTCCTGAGTAAGAA -3'
*Tru*9I -GAC:	5'- GATGAGTCCTGAGTAAGAC -3'
*Tru*9I -GAG	5'- GATGAGTCCTGAGTAAGAG -3'
*Tru*9I -GAT	5'- GATGAGTCCTGAGTAAGAT -3'
*Tru*9I -GTA:	5'- GATGAGTCCTGAGTAAGTA -3'
*Tru*9I -GTC:	5'- GATGAGTCCTGAGTAAGTC -3'
*Tru*9I -GTG	5'- GATGAGTCCTGAGTAAGTG -3'
*Tru*9I -GTT:	5'- GATGAGTCCTGAGTAAGTT -3'
*Eco*RI Selective primers**	
*Eco*RI AA	5'- GACTGCGTACCAATTCAA -3'

Eco RI AT	5'- GACTGCGTACCAATTCAT -3'
Eco RI TA	5'- GACTGCGTACCAATTCTA -3'
Eco RI TT	5'- GACTGCGTACCAATTCTT -3
Eco RI AC	5'- GACTGCGTACCAATTCAC -3'
Eco RI AG	5'- GACTGCGTACCAATTCAG -3'
*Eco*RI TG:	5'- GACTGCGTACCAATTCTG -3'
Eco RI TC	5'- GACTGCGTACCAATTCTC –3'
Eco RI CTG	5'- GACTGCGTACCAATTCCTG -3'
Ec RI GAC	5'- GACTGCGTACCAATTCGAC -3'
Eco RI GAA	5'- GACTGCGTACCAATTCGAA -3'
Eco RI CTA	5'- GACTGCGTACCAATTCCTA -3'
Eco RI AAC	5'- GACTGCGTACCAATTCAAC-3'
Eco RI AAG	5'- GACTGCGTACCAATTCAAG-3'
Ec RI ACA	5'- GACTGCGTACCAATTCACA-3'
Eco RI ACC	5'- GACTGCGTACCAATTCACC-3'
Eco RI ACG	5'- GACTGCGTACCAATTCACG-3'
Eco RI ACT	5'- GACTGCGTACCAATTCACT-3'
Ec RI AGC	5'- GACTGCGTACCAATTCAGC-3'
Eco RI AGG	5'- GACTGCGTACCAATTCAGG-3'
Eco RI GAT	5'- GACTGCGTACCAATTCGAT -3'
Ec RI GAG	5'- GACTGCGTACCAATTCGAG -3'
Ec RI CTT	5'- GACTGCGTACCAATTCCTT -3'
Eco RI CTC	5'- GACTGCGTACCAATTCCTC -3'

**The same PCR primers are used for both the silver stained PAGE and automated DNA analyser options except that for the latter, primers labelled with either HEX or FAM fluorescent dye are used.

9.6. References

Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and M. Zabeau**,** 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. **23**(21): 4407-4414.

10. REMAP & IRAP

REMAP definition: *Any difference in DNA sequence between two genomes, detected by polymerase chain reaction-mediated amplification of the region between a long terminal repeat of a retrotransposon and a nearby microsatellite (Kahl, 2001).*

The dispersion, ubiquity and prevalence of retrotransposon-like elements in plant genomes can be exploited for DNA-fingerprinting. Two DNA techniques based on retrotransposon-like elements are introduced here: IRAP and REMAP (Kalendar et al., 1999). The **IRAP** (**I**nter-**R**etrotransposon **A**mplified **P**olymorphism) markers are generated by the proximity of two retrotransposons using outward facing primers annealing to their long terminal repeats (LTRs). In REMAP (**RE**trotransposon-Microsatellite **A**mplified **P**olymorphism) the DNA sequences between the LTRs and adjacent microsatellites (SSRs) are amplified using appropriate primers.

The principle of IRAP und REMAP is shown in Figure 9.1 below:

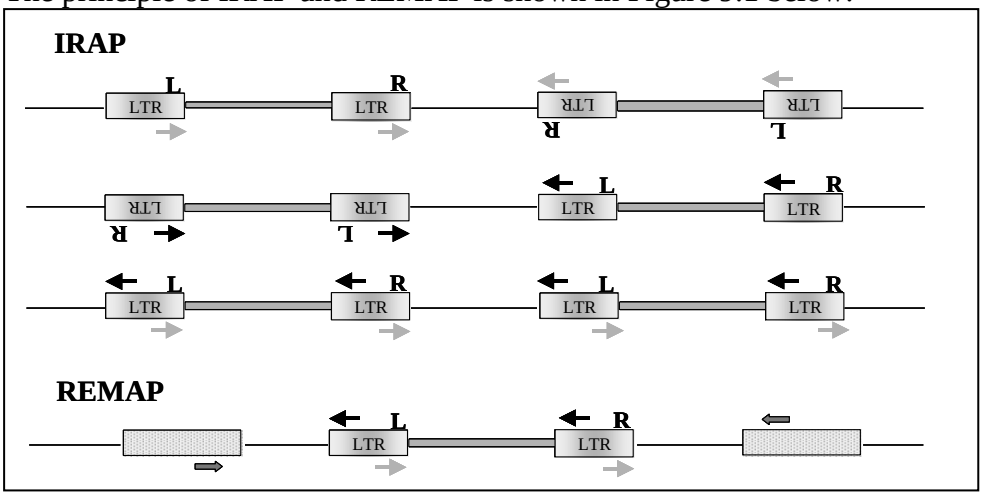


Figure 10-1. Principle of the IRAP und REMAP strategy. **IRAP:** PCR primers facing outward from the 5' (black arrows) and 3' (grey arrows) ends of LTRs will amplify intervening DNA from the retrotransposon in any of the three possible orientations (tail-to-tail, head-to-head, head-to-tail). **REMAP:** LTR primers are used together with a primer consisting of simple sequence repeats (blank boxes) (Kalendar *et al.*, 1999)~~(Kalendar *et al.*, 1999)~~

10.1. Protocol

REMAP and IRAP markers are species specific. In the FAO/IAEA course the following primers for rice and barley were available and used in conjunction with rice and barley DNA.

Table 10.1. LTR primers from the rice retrotransposon Tos17 (Hirochika *et al.*, 1996)~~(Hirochika *et al.*, 1996)~~, sequence and PCR annealing temperatures (T_a).

Primer Sequence T_a

TOS17LTR-1 (outward 3' end of LTR)	TTGGATCTTGTATCTTGTATATAC	56°C
TOS17LTR-2 (outward 3' end of LTR)	GCTAATACTATTGTTAGGTTGCAA	56°C
TOS17LTR-3 (outward 5' end of LTR)	CCAATGGACTGGACATCCGATGGG	56°C
TOS17LTR-4 (outward 5' end of LTR)	CTGGACATGGGCCAACTATACAGT	56°C

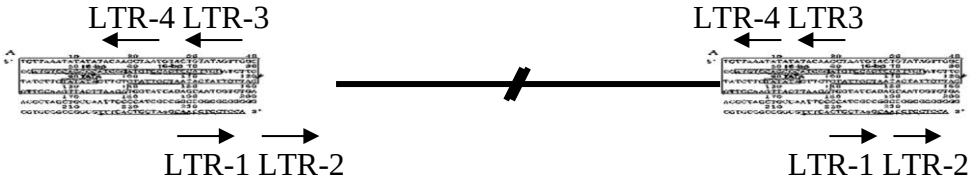
Table 10.2. LTR primers from the barley BARE-1 (Kalendar *et al.*, 1999)~~(Kalendar *et al.*, 1999)~~, sequence and PCR annealing temperatures (T_a).

Primer	Sequence	T_a
BARLTR-2(LTR forward) - IRAP	CTCGCTCGCCCACTACATCAACCGCGTTT ATT	60°C
BARLTR-3(LTR reverse) – IRAP/REMAP	GGAATTCATAGCATGGATAATAAACGAT TATC	60°C

Table 10.3. Microsatellite (SSR) primers and PCR annealing temperatures (T_a).

Sequence	T_a
$(GA)_9C$; $(CT)_9G$; $(CA)_{10}G$	54°C
$(CAC)_7G$; $(GTG)_7C$; $(CAC)_7T$; $GT(CAC)_7$	58°C

NOTE: It is very important to try different combinations of LTR- and microsatellite (SSR) primers for REMAP and LTR-primers for IRAP. Choose primers that have been derived from the species you are working with. The figure below shows you the orientation of only the TOS17-LTR-primers:



NOTE: Gloves and lab coat should be worn throughout.

10.1.1. Prepare a 50µl reaction mix

1. Take a sterile PCR tube and add:

10 x *Taq* buffer	5.0 µl
dNTPs (10 mM)	1.0 µl
Primer 1 (100 pmol/µl)	0.5 µl
Primer 2 (100 pmol/µl)	0.5 µl
DNA (100 ng/µl)	1.0 µl
Taq DNA polymerase (5 U/µl)	0.5 µl
Add ddH_2O to bring volume to	**50 µl**

2. Mix by tapping against the tube.
3. Centrifuge briefly (14,000 rpm for 5 seconds).

10.1.2. PCR amplification

The PCR amplification programme used for the Tos17 sequence was:

Step 1	Initial denaturation	94°C	2 minutes
Step 2	Denaturation	94°C	30 seconds
Step 3	Primer annealing*	T_a	30 seconds
Step 4	Ramp	0.5°C per second to 72°C	
Step 5	Primer extension	72°C	2 minutes
Step 6	Cycling	repeat steps 2-5 for 29 cycles	
Step 7	Final extension	72°C	8 minutes
Step 8	Hold	4°C	forever

* See tables above for appropriate annealing temperatures (T_a).

10.1.3. Separation and visualization of the amplification products

1. Place 15 µl of PCR into a fresh Eppendorf tube.
2. Add 3 µl of 5 X loading buffer containing dye.
3. Vortex briefly.
4. Centrifuge briefly (14,000 rpm for 5 seconds).
5. Load sample into a 2% NuSieve® agarose gel.

NOTE: NuSieve® agarose provides a good separation gel.

6. Run gel for approximately 80 minutes at 80 W (power limiting) or until dark blue front has run 2/3 down the gel.

NOTE: See Section 1 of RFLP Protocol (Agarose gel electrophoresis) for details of gel preparation and running.

7. Stain gel with ethidium bromide (*Caution: ethidium bromide is toxic wear gloves and avoid inhalation*).
8. Visualise bands under UV light (*Caution: wear UV protective glasses and shield your face when you are exposed to the UV light of the transilluminator*).

10.2. References

Hirochika, H., K. Sugimoto, Y. Otsuki, H. Tsugawa, and M. Kanda**,** 1996. Retrotransposons of rice involved in mutations induced by tissue culture. Proc.Natl.Acad.Sci.USA. **93**: 7783-7788

Kalendar, R., T. Grob, A. Regina, A. Suoniemi, and A. Schulman**,** 1999. IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor.Appl.Genet. **98**: 704-711.

10.3. Reagents needed

Use only sterile distilled water for all solutions.

- *Taq* buffer
- dNTPs
- Primers
- *Taq* DNA polymerase (5U/µl)
- DNA (10-20 ng/µl)
- 10 x loading buffer:
  - Glycerol (80%) 600 µl
  - Xylene cyanol 2.5 mg
  - Bromophenol blue 2.5 mg
  - Water 400 µl
- 5 x loading buffer
  - Glycerol (80%) 300 µl
  - Xylene cyanol 1.3 mg
  - Bromophenol blue 1.3 mg
  - Water 400 µl
- Ethidium bromide
- Agarose
- Acrylamide
- Bis-acrylamide
- TBE

11. SINGLE NUCLEOTIDE POLYMORPHISMS (SNPS)

***SNP definition**: Any polymorphism between two genomes that is based on a single nucleotide exchange, small deletion or insertion. (Kahl, 2001).*

Small nucleotide polymorphism (SNP) is a relatively new marker technology originally developed in human. SNPs are the most abundant polymorphic marker with 2 – 3 polymorphic sites every kilobase (Cooper et al., 1985). Originally discovered in humans, SNPs have now been developed for genotyping in plants. SNP technology is heavily dependent upon sequence data. Several methods are available for SNP detection including automated fluorescent sequencing denaturing high-performance liquid chromatography (DHPLC, Underhill *et al*., 1996), DNA microarrays (Hacia and Collins, 1999), single-strand conformational polymorphism-capillary electrophoresis (SSCP-CE, Ren, 2001; Figure 1), microplate-array diagonal-gel electrophoresis (MADGE, Day *et al*., 1998) and matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF, Griffin and Smith, 2000).

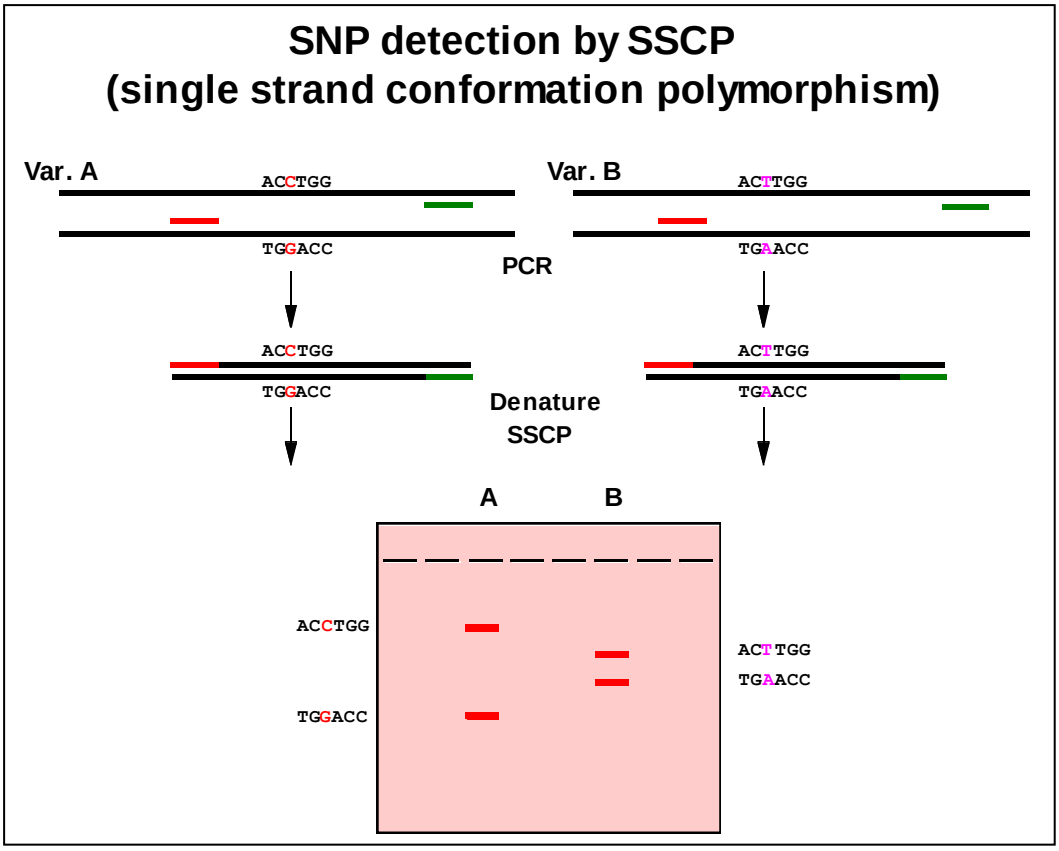


Figure 11-1. The scheme above shows how SNP variation can be detected between varieties A and B (with permission K. Devos).

11.1. References

Cooper, D. N., B. A. Smith, H. J. Cooke, S. Niemann, and J. Schmidtke**,** 1985. An estimate of unique DNA sequence heterozygosity in the human genome. Hum.Genet. **69**(3): 201-205

Day, I. N., E. Spanakis, D Palamand, G. P. Weavind, and S. D. O'Dell**,** 1998. Microplate-arrays diagonal-gel electrophoresis (DADGE) and melt-MADGE: tool for molecular genetic epidemiology. Trends in Biotech. **16**: 287-290

Griffin, T. J. and L. M. Smith**,** 2000. Single-nucleotide polymorphism analysis by MALDI-TOF mass spectrometry. Trends in Biotech. **18**: 77-84

Hacia, J. G. and F. S. Collins**,** 1999. Mutational analysis using oligonucleotide microarrays. J.Med.Genet. **36**: 730-736

Kahl, G., 2001. The Dictionary of Gene Technology. Wiley-VCH, Weinheim.

Ren, J.**,** 2001. High-throughput single-strand conformation polymorphism analysis by capillary electrophoresis. J.Chromatography B.Biomed.Science Appl. **741**: 115-128

Underhill, P. A., L. Jin, R Zemans, P. J. Oefner, and L. L. Cavalli-Sforza**,** 1996. A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc.Natl.Acad.Sci.USA. **93**: 196-200.

12. TILLING

TILLING (Targeting Induced Local Lesions IN Genomes) is a general strategy for the discovery of induced point mutations (COLBERT *et al.* 2001; MCCALLUM *et al.* 2000). The procedure consists of: setting up and running PCR using gene specific primers, denaturing and annealing PCR products to create heteroduplexes between mutant and wild-type strands, digesting heteroduplexes with a single-strand specific nuclease, purifying the products and reducing sample volume, loading sample onto a membrane comb, running the samples on a gel and processing and examining the gel images to identify mutations. The same methods can be used to identify naturally occurring polymorphisms in populations, called Ecotilling, (COMAI *et al.* 2004).

For this training course, we will be using primers for the Arabidopsis OXI1 gene and eight genomic DNA samples, each containing a unique single nucleotide point mutation. The protocol has been scaled down from the standard high throughput TILLING protocol for the discovery of mutations in a large number of pooled samples (TILL *et al.* 2003; TILL *et al.* 2006). Primers and genomic DNA samples are described in a publication on the use of single-strand specific nucleases for mismatch cleavage (TILL *et al.* 2004a). The standard high-throughput TILLING protocol will be followed using fluorescently labelled primers and a LI-COR DNA analyser. Additionally, students will analyse mutations using lower cost and lower throughput agarose gels (for examples see (GALEANO *et al.* 2009; GARVIN and GHARRETT 2007; SATO *et al.* 2006)). The goal of this section of the training course is to familiarize you with the bench and computational techniques that have been developed for TILLING. The hope is that students will leave with a firm understanding of TILLING and the ability to critically evaluate the usefulness of TILLING in his or her research program.

12.1. Protocol

Each group will receive a box containing samples, buffers and solutions for this section of the course. All materials are provided in the box except Ex-Taq polymerase. This will be distributed by the instructor.

12.1.1. PCR reaction with IRDye-labeled primers

Make the following PCR master mix on ice:

72 µl	Water
11.4 µl	10x PCR buffer
13.6 µl	25 mM $MgCl_2$
18.4 µl	2.5 mM each dNTP
8.0 µl	primer cocktail *
1.2 µl	Ex-Taq hot start version

Add 10 µl of PCR mix to each DNA sample (10 µl). Mix sample by pipetting up and down three times.

Place your set of 8 samples in the thermal cycler. Once all teams have deposited their samples, run the PCR cycling program (titled PCRTM70.cyc):

Step 1	Initial denaturation	95°C	2 minutes
Step 2	Denaturation	94°C	20 seconds
Step 3	Primer annealing	73°C (-1°C/cycle)	30 seconds
Step 4	Ramp	0.5°C per second to 72°C	
Step 5	Primer extension	72°C	1 minute
Step 6	Cycling	repeat steps 2-5 for 7 cycles	
Step 7	Denaturation	94°C	20 seconds
Step 8	Primer annealing	65°C	30 seconds
Step 9	Ramp	0.5°C per second to 72°C	
Step 10	Primer extension	72°C	1 minute
Step 11	Cycling	repeat steps 7-10 for 44 cycles	
Step 12	Final extension	72°C	5 minutes
Step 13	Denaturation	99°C	10 minutes
Step 14	Cooling	72°C	20 seconds
Step 15	Cycling	repeat step 14 for 70 cycles (-0.3°C/ cycle)	
Step 16	Hold	4°C	forever

NOTES: For purposes of training, we increase the volume of the master mix so that you have more than is needed. Normally this is not done, but the excess volume controls for pipetting errors and if one group makes a mistake, excess from the other groups can be provided to them.

* The primer cocktail was made in advance as follows:

3 µl forward primer labeled with IRD700 dye (100µM)
2 µl unlabeled forward primer (100µM)
4 µl reverse primer labeled with IRD800 dye (100µM)
1 µl unlabeled reverse primer (100µM)

This mix was stored at -80°C. Prior to use, the mix is thawed on ice, diluted 1:10 with TE (10 mM Tris-HCl, 1 mM ethylene diamine tetraacetic acid (EDTA), pH 7.4) and distributed to each team.

Remove 4µl of samples #7 and #8 and put into new tubes for analysis of PCR product on agarose gel (Step 12.1.3).

12.1.2. Heteroduplex digestion, preparation of Sephadex spin plates

Heteroduplex digestion

Add 4 µl of water to samples #7 and #8 to bring the volume back to 10 µl. Because DNA has been removed for the agarose gel test, these samples should appear weaker on the LI-COR gel.

Prepare the following mix on ice:

Water	326µl
10X CEL I TILLING buffer	60µl
CJE nuclease#	14µl

NOTES:
*10X CEL I buffer is:
5 ml 1M $MgSO_4$
100 µl 10% Triton X-100
5 ml 1M Hepes pH 7.5
5 µl 20 mg/ml bovine serum albumen
2.5 ml 2M KCl
37.5 ml water

The amount of enzyme required will vary depending on nuclease source or possibly from batch to batch of the same enzyme from the same source.

Mix components on ice. Add 40µl of mix to the PCR product and mix by pipetting 2-3 times. Incubate at 45°C for 15 min (in thermal cycler). Cool to 8°C and stop reaction by adding 10 µl of 0.25M EDTA to each sample.

Label a new 8-strip of PCR tubes a set 2 and transfer 35 µl of samples to these tubes. Divide samples by transferring into a new set of 8-tube strip. Set one will be used in Step 12.1.3.1 onwards.

Preparation of Sephadex spin plates

Prior to loading nuclease digested samples onto the denaturing polyacrylamide gel, salts must be separated from the DNA and sample volume reduced to 1.5 µl. There are several methods that can be used to accomplish this. The one you might be most familiar with is alcohol precipitation. For TILLING, we use a different method: size exclusion chromatography using Sephadex G50 medium beads. This is much faster than alcohol precipitation and provides consistent and high recovery of DNA. 96-well plates containing hydrated Sephadex can be prepared up to one week in advance.

Each team will practice preparing a Sephadex plate during the 90°C incubation in Step 12.1.3.1. Pour dry G50 (medium) powder into a 96-hole metal plate and distribute evenly using plastic scraper. Fit a 96-well membrane plate on top, then invert and tap to fill wells with powder. Use a multichannel pipette to add 300 µl water to the top of each well to hydrate, then cover and let sit at least 1 hr at room temperature. Plates are usually made in advance and stored at 4°C in a moist environment for up to one week.

12.1.3. Agarose gel analysis of enzymatic mismatch cleavage, and sample purification

Agarose gel analysis

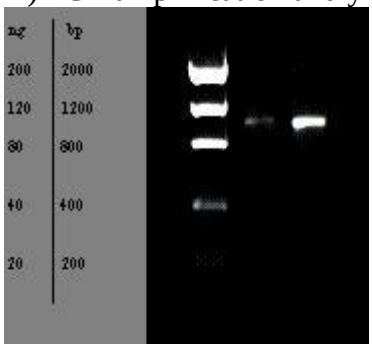
DNA samples are electrophoresed through an agarose gel to verify that (a) PCR was successful in Step 12.1.1 and (b) digestion of mutant DNA by CELI has occurred in Step 12.1.2.

Load samples in the following order:

Lane	1	2	3	4	5	6	7	8	9	10	11
Sample	Low DNA mass ladder	#7 from section 3.1	#8 from section 3.1	#1 from strip 2, section 3.2	#2	#3	#4	#5	#6	#7	#8
Volume (µl)	4	4	4	10	10	10	10	10	10	10	10

Data analysis

A) PCR amplification and yield



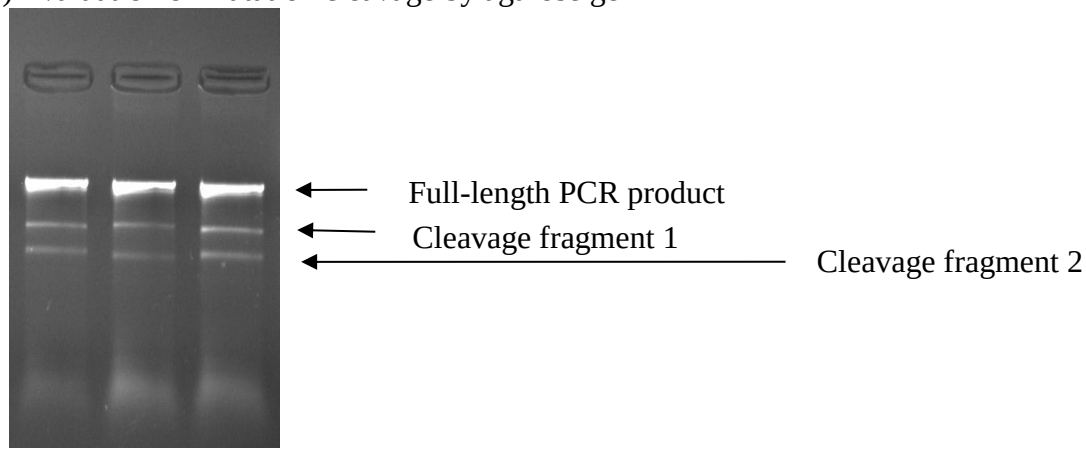
The figure above shows example data of what your first three gel lanes should look like. You should see a single band of the correct size (992 bp). The yield should be at least 7-10 ng/µl of PCR product. The Invitrogen low DNA mass ladder is quantitative and yields are determined

by estimating the intensity of amplified PCR products. For example the intensity of the band in the first PCR sample is between 40 and 20 ng, so the concentration is 30 ng/4 µl or 7.5 ng/µl. The second sample is around 25 ng/µl. Both samples indicate that PCR yield is sufficiently robust for TILLING.

NOTES:

Primer yields are typically not assayed before CEL I digestion of samples. This is done here to evaluate your work. The PBGL typically performs PCR amplification tests on all gene-specific primers prior to purchasing expensive fluorescently labelled primers. Primers passing standardized quality control tests almost always perform well in TILLING experiments.

B) Evaluation of mutation cleavage by agarose gel



DNA used for PCR amplification of samples 1-8 each contains a single point mutation. Cleavage of the mutation creates two fragments of lower molecular weight that migrate faster than the full-length PCR product on the agarose gel. The size of these two fragments equals the size of the full-length PCR product. The eight samples have mutations at different positions on the PCR fragment and so will produce different sized fragments. Take some time to determine where you think mutations are based on the size of your bands.

12.1.4. Sample purification and volume reduction

All of the workshop samples will be loaded onto a single Sephadex plate.
Visually check the Sephadex plate for moistness, and also check underneath for loose Sephadex. If there is any, lightly wipe the bottom with a wet paper towel and gently rinse the

bottom holding the plate on its side. Assemble Sephadex plate, blue plate adaptor, and 96-well skirted 0.2 ml plate (this plate is the "waste" plate).

Spin 2 min at 440g.

Replace the waste plate with a sample catch plate containing 1.5 µl formamide load dye* and 2 µl 200bp marker† in row D. Transfer the entire CEL I reaction sample to each spin plate well. Use a 20-200 µl 8-channel multi-pipettor. Caution: Be sure to dispense liquid to the middle of each well in the Sephadex spin plate, and do not touch the surface of the Sephadex.

Spin 2 min at 440g.

NOTES:
* Formamide load dye is:

250 ml deionized formamide
5 ml 0.5 M EDTA pH 8
60 mg bromophenol blue

† 200 bp marker is made by PCR using gene specific IRD labeled primers that amplify a 200 bp target region. Perform PCR and Sephadex purification as outlined in this protocol. Dilute product to 0.5ng/µl in TE.

The instructor will re-array samples so that all eight samples from a group are adjacent on the LI-COR gel.

Incubate samples at 90°C for approx. 45 min until volume reduced to 1.5 µl.

12.1.5. Preparing, loading, and running LI-COR gels

All student samples will be run on a single gel. The instructor will demonstrate gel preparation.

Clean and assemble glass plates. Prepare the following mixture:

20 ml acrylamide gel mix (6.5%)
15 µl TEMED
150 µl fresh 10% ammonium persulfate

Fill a 20 ml syringe with acrylamide solution. Dispense along the top, avoiding bubbles by rapping just above the liquid edge whenever it appears one might get trapped. If any bubbles

appear, remove them quickly after the gel is poured with a thin wire tool. Leaving a little excess at the well, insert the top spacer all the way and centered. Insert the Plexiglas pressure plate between the glass plate and casting rails. Tighten the top screws as soon the spacer is inserted, compressing the rubber pads on the pressure plate a little. Add acrylamide to the top glass edge where the comb is inserted and on the edges to assure that polymerization is not inhibited within the gel. Let the gel set at least 30 min before putting it into the gel box. Gels can be poured in advance and stored wrapped in a damp paper towel at 4°C for several days.

Loading samples onto membrane combs

All samples will be loaded onto a single loading tray. Each team will load 0.25 µl of sample into the membrane comb loading tray. The instructor will dip the comb into the tray to absorb the sample. The sample should run 1/2 to 2/3 up the length of the comb.

NOTES:

Membrane combs are expensive. To reduce the costs, combs can be reused many times. After the comb has been used, rinse thoroughly with deionized water, soak in water for at least 30 minutes, and allow to dry completely before reuse.

Running LI-COR gels

Pre-run gel 20 min. Gel settings: 1500 V, 40 mA, 40 W, Temp = 50°C, Width = 1028, Speed =2, Channels= 700 & 800

Make sure the back plate is clean and clear of any scratches in the data collection window. Check that the machine is properly focused before loading samples.

Clean the gel slot out with a syringe and drain the top buffer reservoir until the level is below the glass edge. Wick out the remaining buffer, first with a paper towel and then with a 6 inch wide strip of Whatman 1 paper, sliding it into the slot left by the spacer. Using a Pipetteman P1000, fill the slot with 1 ml of 1% Ficoll leaving just a thin bead, ~1 mm above the slot. Hold the comb at a 45°C vertical angle with lane 1 on the left, aim for the slot and insert rapidly by pushing gently until it just touches the gel surface along its length. Gently fill the reservoir to the fill line, insert the electrode/cover, close the top and then click on "Collect image". From the time the comb touches the slot until the time the current is applied should be no more than about 20 min or so to prevent diffusion. After 10 min, open the LI-COR (be sure that you hear the 'pling' signal and the high voltage light goes off), remove the comb and gently rinse the slot with buffer. Replace the upper electrode, close the door and resume the run for 3hrs 45min.

12.1.6. Data Analysis

This component of the TILLING exercise is intended to be performed by students on computers with internet access. Programs and training files along with the protocol below can be downloaded here: http://tilling.fhcrc.org/tillingdemo/computational_tools.shtml.

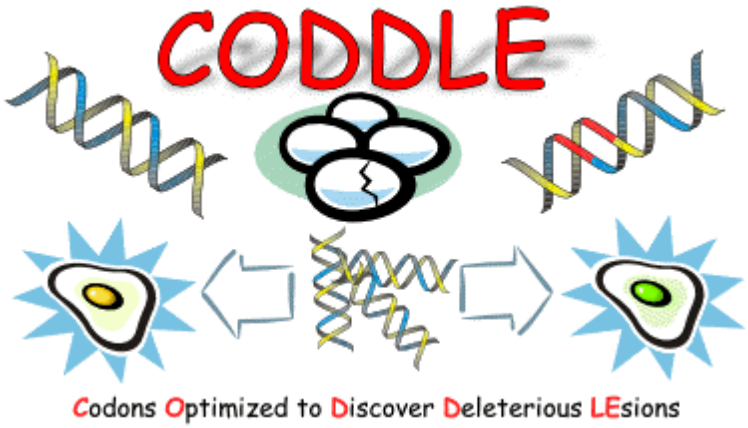
By following the instructions on the webpage, you can easily access all the links described in the protocol below.

12.2. Computation tools

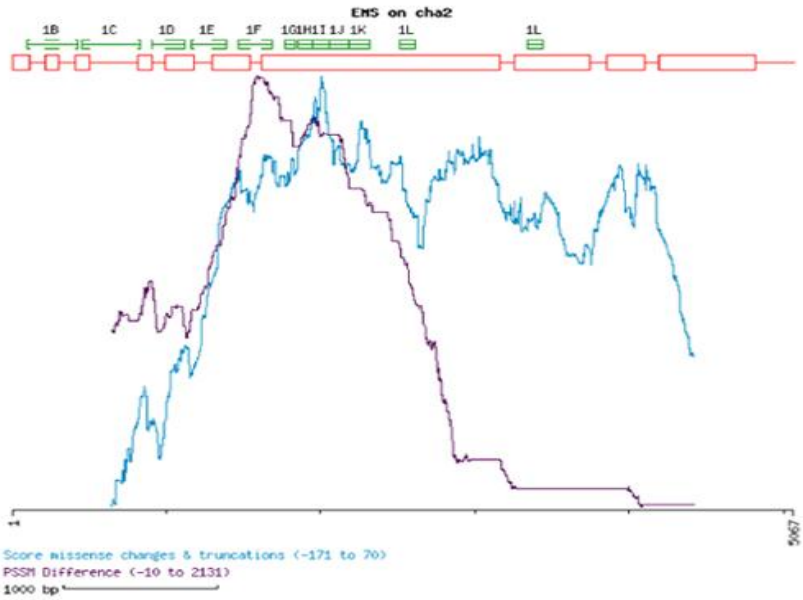
12.2.1. Selecting the best region to screen and designing primers

The current PCR target size for TILLING is between 725 and 1600 bp, with the optimum being around 1.5 kbp. The average gene size in Arabidopsis is 3-4 kb and thus a single PCR amplicon will not cover a whole gene. For genes larger than 1.6 kb, one can either screen the entire gene with overlapping primer pairs (TILLING by tiling), or one can choose the region of a gene with the highest number of possible deleterious changes. For projects where there are a large number of targets, or where the cost of screening could become prohibitive, choosing a "best" screening region is a good approach. This is the approach that STP takes for its public services. For this section of the course, students will use computational tools to choose a target region for TILLING, design primers, and place an order with STP. There are three important components necessary for the optimal TILLING order: 1) a good gene model (intron/exon positions), 2) a good protein sequence homology model, and 3) a good PCR primer pair.

These choices are facilitated by the CODDLE Input Utility, (http://www.proweb.org/input/) which accepts genomic, cDNA and/or protein sequences from your own files or via links from public databases.



1. Open the **Test Genes page** (http://tilling.fhcrc.org/tillingdemo/CODDLEtestgenes/) in a new browser window. Select a gene by clicking on the gene name. Here you will find both genomic and protein sequence information.
2. Select and copy the genomic DNA sequence.
3. Open **CODDLE Input Utility** (http://www.proweb.org/input/) in a new browser window.
4. In the CODDLE input page, enter the gene name and paste in the genomic sequence information.
5. Go back to the gene page and copy the protein sequence.
6. Paste the protein sequence in the appropriate window.
7. Click the "Begin Processing" button. The CODDLE input utility is now creating a gene model and searching for homology information that will help identify regions that are likely to be important for protein function.
8. A new window should appear with a summary of the Blocks family protein homology, an intron/exon join statement and the amino acid sequence. Click the "Proceed with CODDLE" button.
9. In the CODDLE page, select "TILLING w/EMS (plants)" as the mutation method, then click "CODDLE your gene". CODDLE will now evaluate every possible mutation and provide a high scoring window where the highest number of deleterious changes are likely to be found. A new window will open with the CODDLE output. The graphical output shows the gene model (red boxes and lines), protein homology (green boxes) and the score of the gene (purple and blue lines). The purple line indicates the score for predicted deleterious missense changes, and the blue line is the score for the total number of non-silent changes. In the example below, the highest scoring window for missense and truncation changes is centred at position 2008.



Below the graph is information on the Blocks protein homology and an additional options box where you can examine a region of the gene that was not selected as the high scoring region. Below this, the changes and predicted effect of the changes can be seen at the sequence level. For a complete description of the symbols used, and more detailed information on CODDLE, please visit the **CODDLE glossary**.

10. When you are satisfied with the CODDLE output, click “Create primers for this window”.
11. Evaluate the information in the Primer3 window. Note that the optimum Tm for primers is 70°C. Click “pick primers”.
12. In the output page, click “display this pair of primers” for your favorite set of primers.
13. You will now be directed to a page summarizing your primer choices. Note that the percentage of each type of change is listed.
14. When satisfied, click “order TILLING of this region”.
15. You are now directed to an STP order page. Enter the following email address: test@fhcrc.org and select Arabidopsis as the organism.
16. Click “place order”. Your order will now be searched in the STP database. If the target has been previously screened, you will be provided with information on found mutations. If it is a new target, it will be blasted against the Arabidopsis genome to ensure that the primers are designed to the correct organism. Once ready, click “store” to store the order in the database.

17. NOTES: The CODDLE input utility, CODDLE and Primer3 are all general tools that are available on the World Wide Web. You may find them useful for non-TILLING applications. Steps 14-16 have been included to illustrate that placing, verifying and confirming orders are tasks that have been automated by STP.

18. Additional Exercises: Once you have familiarized yourself with CODDLE and primer design, try inputting other information in the CODDLE input utility such as the Genbank URL of your favorite sequence (step 4). Also, try making additional Blocks with the SIFT programme (step 8). Finally, use the additional options window of the CODDLE output (step 9) to design primers to a different region of the gene.

12.3. Data analysis

The programme GelBuddy has been created to assist the discovery of mutations and polymorphisms ((ZERR and HENIKOFF 2005). It is available as a free download (http://www.proweb.org/gelbuddy/). This program should already be loaded onto the training course computers. For this exercise, download sample images from here (http://tilling.fhcrc.org/tillingdemo/ImagesforFAOgelBud/). Be sure to download both the IRD700 and IRD800 images. The protocol below uses the basic Gelbuddy features for analysis of a standard TILLING gel. Tools are provided for the analysis of EcoTILLING or two dimensionally pooled gels that are not described. More information can be found at the GelBuddy page.

1. Download IRD700 and 800 jpeg or tiff images to your desktop. For example, download both 43ugfp115a_bt.7 and 43ugfp115a_bt.8.
2. Open Gel Buddy.
3. Import images. Under file, choose “Open 700 and 800 channel images”.
4. Select the first image to load. While holding down the shift key, select the second image. Click “open”.

5. Adjust the 700 channel image to the desired intensity using the slider bars located on the upper region of the GelBuddy window.

6. Adjust the 800 channel image. Click the 700-800 box at the top of the window to switch to the 800 channel. With the 800 channel selected, adjust the image as in step 5.

7. Call lanes. Click the "find lanes" box located in the tool bar at the top of the window.

8. Set the number of sample lanes in the "find lanes" pop up window (the default is 96 for a standard TILLING run). Select segmented lane tracks. Unless one of the channels is very bad, use the both channels for detecting lanes. Click "ok".
9. Editing lanes. The blue lane markers should run through the lanes with the 200 bp marker. If they do not, or one or more lanes are called wrong, click the "edit lanes mode" in the toolbar.

10. Select the lane you wish to edit or the lane adjacent to the area where you wish to add a lane. Under the edit menu, select insert or delete lanes as required. If a lane merely needs to be "straightened", select the boxed regions and drag to the desired location.
11. Click the "show lanes box" to remove lines.
12. Set the molecular weight migration. Click the "show calibration information" box. Vertical lines will appear.

13. Place the mouse over one of the numbers in blue and drag that number to the desired location on the gel. The 700 should align with the highest band in the ladder lanes. The 200 should align with the 200 bp marker.
14. Now set the 0% and 100% migration by dragging the red numbers to the bottom of the signal on the gel image (100%) and to the top of the full length product (0%). When complete, click the "calibration information" box again to make lines disappear.
15. Select mutations. Select the "record signals mode" box. Using the 700-800 box, switch between channels to find mutations. You will be prompted to enter the size of the full length product (0% migration). Enter the number at 0% and click "ok". Enter your initials in the "created by" box. The signal grouping should be set to "all lanes". Click the mouse over the mutation to select the mutation. When selecting mutations, note that mutations in the 700 channel are marked red and those in the 800 are marked with a blue box. If you are unsure of a mutation, note that the size of the band is given at the bottom of the window when your mouse is over the mutation. For any one lane, the sizes of bands

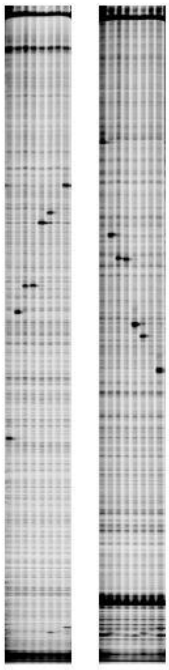
in the blue and red boxes should equal the full length product. Do not be alarmed if the sizes are up to 100 base pairs off. To delete a box, hold down the option key and click the box.

16. Once you have selected all of the mutations, select the "show signals" box to remove the boxes. Look at the gel again to be sure you have selected real mutations. Select the box again to make the boxes reappear.

17. To zoom in to a region of the gel, select the "zoom in mode" box and click on the region you wish to enlarge. To zoom out, select the "zoom out mode" box. To fit the image back to the original window, select the "zoom to window" box.

18. When you have finished analysing the gel, click the log box to see a report. Inspect the signals sorted by lane table. True mutations should have paired signals in the 700 and 800 channel that add up to the full-length product size.

19. Compare your data with what was found by STP. At STP, data from GelBuddy is directly posted to the program Squint in the STP database using the GelBuddy autopost function. You can view squint files for this exercise here (http://tilling.fhcrc.org/cgi-bin/displayWorkshop.pl?form=newSquint). Under "squinting", click "new/modify/view". In the LI-COR run name field, enter the run name. The run name does not include .7.jpg or .8.jpg. For instance, for the first set of images on the images page, you would type 42600m1a_eb as the run name. Select "list current squint file" in the select a squint action box. Click the submit button to view the squint file. Did you find all the mutations? Did you find more than were reported? Note that mutations are given a confidence score based on quality. Confidence level A: the bands in both channels are clear and add up to the full-length size; level B: there are two corresponding bands but one of the bands is questionable; level C: data is available for one of the two channels but the band is most likely a mutation; level D: data is available for one of the two channels and the band is weak.
20. Try some other features in GelBuddy. For weak bands, try the "show inverted image" box to view the inverted image. Click the calibration box to show the horizontal calibration lines. Under the options pull down menu, try changing some of the calibration settings and see what happens to the lines. Notice that GelBuddy is compensating for lane to lane variation such as gel smiling. Want to see what the samples you processed should look like? Below is a test gel of these samples run in Seattle.



Mutations in the Arabidopsis OXI1 gene. IRD700 left, IRD800 right. Lanes 1-8 equal samples 1-8 in the training course.

Lane	Fragment size in bp (IRD700)
1	288
2	441
3	476
4	477
5	566
6	580
7	603 *
8	624

*The mutation in lane 7 is homozygous and thus is not detected when screened alone. For the course, an equal amount of wild-type DNA has been added to sample 7 so that hetereoduplexes will be formed.

12.4. Additional info

12.4.1. List of consumables and equipment

Note that not all equipment is necessary for a successful TILLING operation, and not all equipment may be available. For instance, the comb-loading robot is no longer being sold by MWG, and neither are the thermal cyclers. Manual comb-loading is relatively easy, and most thermal cyclers should work for TILLING, so lower cost options are available.

Lab Supplies

Product	Company	Catalog Number
LI-COR 4300 S DNA analyzer	LI-COR	4300-02
Apricot pipettor	Perkin Elmer	PP-550
Combloader	MWG	Combload
Centrifuge 5804 (Cel I)	Brinkman	2262250-1
Thermocycler Primus 96	MWG	4000-000005
Centrifuge 5810 (Genomic)	Brinkman	2262500-4
Nanopure (Water Treatment)	VWR (Barnstead)	13500-866
Centrifuge 5417C (PCR bench)	Brinkman	2262170-0

Equalizer (electric pipettes)	Matrix	2139
Heat plate sealer	Marsh (AB Gene)	AB-0384
pH meter	Fisher	13-636-AR10
Heat blocks	VWR	52434-232
Pipettes LTS multi-channel 20-1000µl	Rainin	L8-20, L8-200…
APC surge protector	CDWG	323633
Multi heat block	Fisher	NC9800611
Pipettes LTS single channel 20-1000µl	Rainin	L-20, L-200…
Stir plate	Fisher	11-500-49SH

Consumables

Product	Company	Catalog Number
Membrane combs	Gel Company	CAJ96
MWG 96-well plates	MWG	4050-000003
QT tip,250ul clear, sterile filter tip	Molecular Bio Products	1043-60-5
QT tip, 500ul clear, sterile non-filter tip	Molecular Bio Products	1043-61-7
Acrylamide	Li Cor	82705607
Buffer reservoirs	Apogent Discoveries	8094
Sephadex G-50	A.Pharmacia	17-0043-02
EDTA	Research Organics	3002E
Ficol	Fisher	BP525-25
Tris	Research Organics	30960T
Boric acid	Research Organics	1748B
Milipore plates	Fisher	MAHVN 4550
Formamide	Sigma	F5786
Sealing tape PCR	Island Scientific	IS-609
Sealing tape non-PCR	Island Scientific	IS-SEAL
IRD 700	LI-COR	4200-60
IRD 800	LI-COR	4000-45
Taq, dNTP, PCR buffer	Pan Vera	TAK RR001C
Seq direct clean-up kit	Qbiogene	9904-200
EZPeel clear heat seal	Marsh Bio Prod	AB-0812
EZPeel aluminium heat seal	Marsh Bio Prod	AB-0745
LTS tips 10F	Rainin	GP-L10F
LTS tips 10S	Rainin	GP-L10S
LTS tips 200F	Rainin	GP-L200F
LTS tips 250S	Rainin	GP-L250S
LTS tips 1000F	Rainin	GP-L1000F
LTS tips 1000S	Rainin	GP-L1000S
20uL LTS tips spacesaver	Rainin	GPS-L10
200uL LTS tips spacesaver	Rainin	GPS-L250S
1000uLLTS tips spacesaver	Rainin	GPS-L1000S

Sephadex column loader 45ul	Fisher	MACL09645
Sephadex scraper replacement	Fisher	MACL0SC03

12.5. Frequently asked questions

Will TILLING work in my favourite organism?

TILLING is a general method and should work for most organisms. Requirements include the ability to induce mutations, propagate and/or store mutant organisms and PCR amplify gene specific targets.

What about polyploids or duplicated gene targets?

STP has successfully screened polyploid species. Additionally, Slade, *et al*., have published TILLING data for polyploid wheat (SLADE *et al*. 2005). For polyploids and duplicated gene targets, a good approach is to pre-test unlabeled primers before purchasing IRD labeled primers. This is the approach taken for the Maize TILLING Project (http://genome.purdue.edu/maizetilling/). Following PCR and agarose gel analysis, products are sequenced. Primer pairs are selected for TILLING if they produce at least 7 ng/µl of product and sequence analysis indicates the amplification of a single target.

What if there is no genomic sequence available for my organism?

Short of cloning genes, you can design primers to EST data (or whatever is available) and pre-screen the primers. Sequencing the PCR products will provide genomic sequence information. It is important to select primers that yield products within the appropriate size range for your assay. Also, you may wish to avoid TILLING large amounts of intron as mutations in introns are likely to be non-functional. You may be able to use genomic sequence from a related organism to guess at the position of introns in your organism.

I do not have access to a LI-COR, can I still TILL?

The choice of read out platform (the machine used), can affect the level of allowable pooling, rate of false positives and negatives, robustness of the assay, as well as other factors. Thus, the choice of read out platform can have a large impact on the cost and throughput of your operation. STP has exclusively used LI-CORs and therefore it is difficult to comment directly on other platforms. Perry et al. published TILLING work using an ABI 377 (PERRY et al. 2003). Other end labeling strategies, such as using radioactivity, should work. Again, the throughput, efficiency and screening cost associated with the platform should be considered.

An alternative to end labeling is body labeling. Body labeling DNA may not be as efficient as end labeling either the DNA or a probe. That said, one can use single-strand specific nucleases to induce double strand breaks in DNA, allowing visualization on native agarose gels (BURDON and LEES 1985; CHAUDHRY and WEINFELD 1995; HOWARD et al. 1999; SOKURENKO *et al*. 2001) Most likely, this will prove to be a lower throughput option.

I am more interested in EcoTILLING. How is it different?

EcoTILLING is a method for the discovery and genotyping of natural polymorphisms (COMAI *et al.* 2004). The starting material for EcoTILLING is DNA from "natural" populations rather than mutagenized ones. Depending on the population, one might expect a substantially higher frequency of polymorphisms than the rare induced mutations found in a chemically mutagenized population. The wet bench protocols used for TILLING and EcoTILLING are the same. GelBuddy has been designed to work with EcoTILLING data and some EcoTILLING-specific features are available in GelBuddy.

Will a chemical mutagen be effective on all genes? What about background mutations in the lines? Do I need a license to TILL?

For answers to these questions, please see the STP FAQ page (http://tilling.fhcrc.org/files/FAQ.html).

12.6. Additional protocols

12.6.1. Sequencing

This protocol is a scaled down version of the standard high-throughput sequencing protocol.

H_2O	54.8µl
Ex Taq buffer	10µl
dNTP	8µl
forward primer (10 µM)	1µl
reverse primer (10 µM)	1µl
HS Ex Taq	0.25µl

Add 15 µl mix to 5 µl DNA and mix well.

Run the following programme:

Step 1	Initial denaturation	95°C	2 minutes
Step 2	Denaturation	94°C	20 seconds
Step 3	Primer annealing	73°C (-1°C/cycle)	30 seconds
Step 4	Ramp	0.5°C per second to 72°C	
Step 5	Primer extension	72°C	1 minute
Step 6	Cycling	repeat steps 2-5 for 7 cycles	
Step 7	Denaturation	94°C	20 seconds
Step 8	Primer annealing	65°C	30 seconds
Step 9	Ramp	0.5°C per second to 72°C	
Step 10	Primer extension	72°C	1 minute
Step 11	Cycling	repeat steps 7-10 for 44 cycles	
Step 12	Final extension	72°C	5 minutes
Step 13	Hold	4°C	forever

Quantify yield on an agarose gel (this is normally done only on 1 row of a 96 well plate).

Pre-sequencing clean-up:

To 10 µl PCR product add and mix well:

*4 µl Shrimp alkaline phosphatase

*1 µl Endonuclease I (keep enzymes on ice at all times)

*Check company protocol for units/µl

Incubate 37°C for 15 min., 80°C for 15 min. (Follow manufacturer's suggestion).

The pre-sequencing amplification is performed with the unlabeled primers used in the TILLING screen. Following the manufacturer's protocol, HS Ex-Taq (Takara) is used in a 20 µl final reaction volume with 0.005 ng genomic DNA (for Arabidopsis).

Sequencing RXN (Big Dye version 3.0 or higher/ ABI 3100 or higher)

Add 5 µl of 5% DMSO to PCR product and mix

To new set of tubes add:

4 µl diluted Big Dye (version 3.0 or higher) (1:1 dilution with PCR H_2O)

1 µl forward primer (3 µM)

5 µl PCR product (diluted with DMSO)

Mix well and spin down.

Step 1	Initial denaturation	95°C	5 minutes
Step 2	Denaturation	95°C (ramp at 1°C/sec)	10 seconds
Step 3	Primer annealing	50°C (ramp at 1°C/sec)	5 seconds
Step 4	Primer extension	60°C (ramp at 1°C/sec)	4 minutes
Step 5	Cycling	repeat steps 2-4 for 24 cycles	
Step 6	Hold	8°C (ramp at 1°C/sec)	forever

Big dye removal and running the ABI is performed by a core facility.

Sequence trace analysis is performed using Sequencher™ 4.5 software (Gene Codes). Both heterozygous and homozygous mutations can be confirmed utilizing the mapping information gathered in the TILLING screens.

12.7. EMS mutagenesis of Arabidopsis seed

EMS mutagenesis of maize pollen for the population used in the Maize TILLING Project has been described (TILL *et al.* 2004b).

12.7.1. Materials

Orbital shaker: Aros 160 with a 1.25 cm radius of gyration
10-15 L tub
Microfuge tubes with 50 mg of seed each
Stir plate and stir bar
1000 ml beaker
1 L 2 N NaOH
Squeeze bottle of di H_2O
10% Tween 20
P-1,000 pipetter with barrier tips. Some of these ought to have notches cut in Tip as per "A Note on Technique."
P-20 pipetter with barrier tips
EMS (methanesulphonic acid ethyl ester), Sigma
Glass scintillation vials I.D. = 2.5 cm
Box for dry hazardous materials disposal
Plastic bag for hazardous materials disposal
Box of nitrile (not latex) gloves and a lab coat

12.7.2. Standard size batch

In order to avoid variation in mutation rate that could arise from scaling properties, the first 10 mutagenesis procedures for this project except the 6th were done in standard batches of 50 mg seed in 4ml of EMS solution. Only flat-bottomed glass scintillation vials of 2.5 cm ID were used so as to avoid subtle variations in the agitation of the seeds. This standard procedure did not make the concentration of EMS a good predictor of the EL count. Because of this, and to allow reducing the number of people needed to care for a batch of M_1 plants, quantities of seeds less than 50 mg are now allowed.

12.7.3. A note on technique

Before beginning this procedure, cut a couple of notches in the tip of several of the P1000 tips. If the notch is too small to allow seeds to pass through, the tip can be pressed against the bottom of the scintillation vial and the supernatant can be efficiently aspirated without loss of seeds.

Day 1:

1. Preparation of Fume Hood for procedure.
   1.1. Label each scintillation vial with the concentration of EMS that is to be used in it.
   1.2. Warn all personnel that a dangerous procedure is about to be performed in the hood.
   1.3. Place all materials in hood.
   1.4. Put 125 ml of 2 N NaOH and 375 ml of H_2O in beaker with stir bar slowly rotating. Place remaining 875 ml of 2 N NaOH in tub with 2.6 L H_2O.
2. Add 4 ml of H_2O to each vial and mark level with a fine tip marker then empty vial of H_2O.
3. Rinse seed into each vial with 4 ml of diH_2O. Add 40 ml of 10% Tween 20 to each vial and agitate at 180 RPM for 15 sec.
4. Pipette off Tween/ H_2O and add 4ml DI-H_2O to each vial. Agitate for 5 min at 180 RPM. Repeat for 4 total washes.
5. Add DI-H_2O to each vial to 4 ml line made in 2) in order to achieve a total volume of 4 ml.
6. Use gloves, lab jacket, and fear for following steps.
7. Add .425 X (ml) EMS to each vial with barrier tip P-20s. X is desired [EMS] (mM). Dispose of tips in beaker of 0.5 N NaOH.
8. Agitate for 17 hr at 180 RPM at room temperature.

Day 2:

1. Pipette off EMS solution from each vial and dispose in flask of 0.5 N NaOH.
2. Fill each vial to shoulder with di H_2O from squeeze bottle, swirl by hand, then pipette off supernatant and dispose as in 1). Repeat 5 times.
3. Add diH_2O to vial to achieve 4 ml and agitate 15 sec.
4. Pipette off as in 2) and repeat.
5. Store at 4°C until sown.
6. Allow NaOH that has been used for EMS disposal to stir for 30 min., then gently pour contents of beaker into tub of 0.5 N NaOH, placing beaker in tub as well, then pour down drain and flush with cold running water for 15 min.
7. Wipe off pipettes and inside of hood with dil NaOH, and call Hazardous Materials Disposal to remove solid waste.

12.7.4. DNA extraction

DNA isolation is done per FastDNA a kit protocol (revision #6540-999-1D04, http://www.qbiogene.com/fastprep/protocols.shtml), with the following variations and warnings:

1. Use only one ceramic bead per shaker tube.
2. Run shaker for 45 min at 4.5 m/s.
3. The first centrifuge spin should be at 14,000 × g for up to 30 min. Draw off as much as 800 - 900 ml supernatant from shaker tube in Step 3 of the FastDNA protocol.
4. After DNA is bound in a pellet to the Binding Matrix, take care not to disturb the pellet when discarding supernatants in Step 4 of the FastDNA protocol.
5. To make re-suspension easier, all spins before a re-suspension (both 1-minute spins in Step 4 of the Fast DNA protocol) should be at 9,000-10,000 × g for 3 min.
6. To re-suspend a pellet (Steps 4 and 5 of the Fast DNA protocol), use the vortex or noisily rake the tube across a tube rack, a practice known as "ducking" for the quack-like sound made. When ducking, take care to hold down the cap of the tube to prevent it from popping open.

In Step 5 of the Fast DNA protocol, elute binding matrix with 200 ml DES. Spin at 14000 × g for ~5 min. Then pipette off 180 ml of supernatant, taking extreme care not to draw up particles of Binding Matrix, and transfer supernatant to a sterile screw-top tube. Add 20 ml of 10x TE @ 3.2 m g/ml RNAse A.

12.8. References

BURDON, M. G., and J. H. LEES, 1985 Double-strand cleavage at a two-base deletion mismatch in a DNA heteroduplex by nuclease S1. Biosci Rep **5:** 627-632.

CHAUDHRY, M. A., and M. WEINFELD, 1995 Induction of double-strand breaks by S1 nuclease, mung bean nuclease and nuclease P1 in DNA containing abasic sites and nicks. Nucleic Acids Res **23:** 3805-3809.

COLBERT, T., B. J. TILL, R. TOMPA, S. REYNOLDS, M. N. STEINE *et al.*, 2001 High-throughput screening for induced point mutations. Plant Physiol **126:** 480-484.

COMAI, L., K. YOUNG, B. J. TILL, S. H. REYNOLDS, E. A. GREENE *et al.*, 2004 Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. Plant J **37:** 778-786.

GALEANO, C. H., M. GOMEZ, L. M. RODRIGUEZ and M. W. BLAIR, 2009 CEL I Nuclease Digestion for SNP Discovery and Marker Development in Common Bean (Phaseolus vulgaris L.). Crop Science **49:** 381-394.

GARVIN, M. R., and A. J. GHARRETT, 2007 DEco-TILLING: an inexpensive method for single nucleotide polymorphism discovery that reduces ascertainment bias. Molecular Ecology Notes **7:** 735-746.

HOWARD, J. T., J. WARD, J. N. WATSON and K. H. ROUX, 1999 Heteroduplex cleavage analysis using S1 nuclease. Biotechniques **27:** 18-19.

MCCALLUM, C. M., L. COMAI, E. A. GREENE and S. HENIKOFF, 2000 Targeted screening for induced mutations. Nat Biotechnol **18:** 455-457.

PERRY, J. A., T. L. WANG, T. J. WELHAM, S. GARDNER, J. M. PIKE *et al.*, 2003 A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume Lotus japonicus. Plant Physiol **131:** 866-871.

SATO, Y., K. SHIRASAWA, Y. TAKAHASHI, M. NISHIMURA and T. NISHIO, 2006 Mutant Selection from Progeny of Gamma-ray-irradiated Rice by DNA Heteroduplex Cleavage using Brassica Petiole Extract. Breeding Science **56:** 179-183.

SLADE, A. J., S. I. FUERSTENBERG, D. LOEFFLER, M. N. STEINE and D. FACCIOTTI, 2005 A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. Nat Biotechnol **23:** 75-81.

SOKURENKO, E. V., V. TCHESNOKOVA, A. T. YEUNG, C. A. OLEYKOWSKI, E. TRINTCHINA *et al.*, 2001 Detection of simple mutations and polymorphisms in large genomic regions. Nucleic Acids Res **29:** E111.

TILL, B. J., C. BURTNER, L. COMAI and S. HENIKOFF, 2004a Mismatch cleavage by single-strand specific nucleases. Nucleic Acids Res **32:** 2632-2641.

TILL, B. J., S. H. REYNOLDS, E. A. GREENE, C. A. CODOMO, L. C. ENNS *et al.*, 2003 Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Res **13:** 524-530.

TILL, B. J., S. H. REYNOLDS, C. WEIL, N. SPRINGER, C. BURTNER *et al.*, 2004b Discovery of induced point mutations in maize genes by TILLING. BMC Plant Biol **4:** 12.

TILL, B. J., T. ZERR, L. COMAI and S. HENIKOFF, 2006 A protocol for TILLING and Ecotilling in plants and animals. Nat Protoc **1:** 2465-2477.

ZERR, T., and S. HENIKOFF, 2005 Automated band mapping in electrophoretic gel images using background information. Nucleic Acids Res **33:** 2806-2812.

13. ALTERNATIVE ENZYMOLOGY FOR MISTMATCH CLEAVAGE FOR TILLING AND ECOTILLING: EXTRACTION OF ENZYMES FROM WEEDY PLANTS

13.1. Objective

A crude celery extract containing the single-strand-specific nuclease CELI has been widely used in TILLING and Ecotilling projects around the world. Yet, celery is hard to come by in some Member States. Based on previous studies and bioinformatic analysis suggestion homologies exist to CELI in all plants. Therefore, we developed a protocol for extraction of active enzyme from plants common across the world: weeds. We isolated weed plants from the grassland around the Seibersdorf laboratories and isolated a crude enzyme extract (in parallel to the enzyme extracts from celery). Since, there was no or only very low mismatch digestion activity in the crude extract, we applied a centrifuge-based filter method to concentrate the enzyme extract.

13.2. Materials

MATERIALS / BUFFERS FOR ENZYME EXTRACTIONS	Notes
hand-held mixer (or juicer)	From any supplier
STOCK: 100mM PMSF (stock in isopropanol)	To prepare an aqueous solution of 100µM PMSF (for buffers A and B), add 1 ml 0.1M PMSF per liter of solution immediately before use.
STOCK: 1M Tris-HCl, pH 7.7.	
Buffer A: 0.1 M Tris-HCl, pH 7.7, 100 µM PMSF.	
Buffer B: 0.1 M Tris-HCl, pH7.7, 0.5 M KCl, 100 µM PMSF.	
Dialysis tubing with a 10,000 Dalton molecular weight cut off (MWCO)	e.g. Spectra/PorR Membrane MWCO: 10,000, Spectrum Laboratories, Inc.
$(NH_4)_2SO4$ (Ammonium sulphate)	
Sorvall Centrifuge	Or equivalent centrifuge/rotor combination to achieve needed gravitational force
MATERIALS FOR CONCENTRATION OF ENZYME EXTRACTS	
Amicon Ultra Centrifugal Filters (0.5mL, 10K)	Millipore Amicon Ref.No. UFC501024 24Pk
Refrigerated (4°C) Microcentrifuge	e.g. Eppendorf 5415R

TILLING-PCR	
Thermocycler	e.g. Biorad C1000 Thermal cycler
PCR tubes	Life Science No 781340
TaKaRa Ex Taq™ Polymerase (5U/ul)	TaKaRA
10X Ex Taq™ Reaction Buffer	TaKaRa
dNTP Mixture (2.5mM of each dNTP)	TaKaRa
Agarose gel equipment	

13.3. Methods

13.3.1. Enzyme extraction

1. Collect approximately 200 grams of mixed monocot and dicot weedy plants were collected that were growing on the periphery of our sorghum field.

2. Wash material 3x in water and then ground using a hand-held mixer and by adding about 300 mls of water to facilitate tissue disruption (or optional in a juicer)

3. Add 1M Tris-HCl (pH7.7) and 100mM PMSF to a final concentration of Buffer A (0.1M Tris-HCl and 100µM PMSF) (NOTE: Stocks and water should be kept at 4°C, perform subsequent steps at 4°C)

4. Centrifuge for 20 min at 2600 x g in Sorvall GSA rotor to pellet debris. Save supernatant.

5. Bring the supernatant to 25% ammonium sulphate (add 144 g per liter of solution). Mix gently at 4^{o}C (cold room) for 30 min.

6. Centrifuge for 40 min at 4°C at ~14,000 x g in sorvall GSA rotor (~9000 rpm). Discard the pellet.

7. Bring the supernatant to 80% ammonium sulphate (add 390 g per liter of solution). Mix gently at 4^{o}C for 30 min.

8. Centrifuge for 1.5 hours at 4°C at ~14,000 x g in sorvall GSA rotor. SAVE the pellet. Discard the supernatant (be careful in decanting the supernatant!) The pellet can be stored at -80^{o}C for at least two weeks.

9. OPTIONAL: Pellets can be frozen at -80°C for months.

10. Resuspend the pellets in ~ 1/10 the starting volume with Buffer B (Frozen pellets of the weed juice extract were suspended in 15mL Buffer B and pellets of the celery juice extract in 10 mL Buffer B). Ensure the pellet is thoroughly resuspended.

11. Dialyze against Buffer B at 4°C (2 Liters per 10mls of resuspended solution).. Use e.g. Spectra por 7 MWCO 10000 tubing. (NOTE: Soak the dialysis tubing in nanopure water for 30 min. before use.)

12. Dialyze for 1 hour against Buffer B at 4°C

13. Repeat for a total of 4 dialysis steps with a minimum of 4 hours dialysys. (NOTE: Longer dialysis is better, it is often convenient to perform the third dialysis overnight).

14. Remove liquid from dialysis tubing. It is convenient to store ~75% of the liquid in a single tube at -80°C and the remainder in small aliquot for testing. This protein mixture does not require storage in glycerol and remains stable through multiple freeze-thaw cycles, however, limiting freeze thaw cycles to 5 limits the chance of reduced enzyme activity

15. Perform activity test (step 3.3, or proceed immediately to enzyme concentration, step 3.2)

Figure 1. Mixture of different plant species (weedy plants) from the grassland around the Seibersdorf laboratories used for the isolation of an enzyme extract for mismatch cleavage.

13.3.2. Concentration of enzyme extractions

Concentration of weed and celery enzyme extracts is done using Amicon Ultra 10K centrifugal filter devices (for 0.5mL starting volume; in 1.5-mL tubes).

1. Perform with 600µL of protein extract after dialysis
2. Clear extract by centrifugation at 30 min / 10,.000 x g / 4°C (to pellet plant material) in refrigerated microcentrifuge

3. Transfer 500 µL of the (cleared) supernatant to a filter device (keep the rest of the supernatant as control "before concentration").

4. Centrifuge the filter device with a collection tube inserted per manufacturer's instructions for 30 min / 14,000 x g / 4°C

5. Remove filter device, invert and place in new collection tube.

6. Centrifuge for 2 min / 1,000 x g / 4°C

7. Measure the recovered volume. This is your concentrated protein. Calculate the concentration factor with the following formula: Starting volume/Final folume = concentration factor

13.3.3. Test of Mismatch Cleavage Activity

1. Produce TILLING-PCR products for mismatch cleavage tests with the concentrated enzyme extracts. The exmple below is for barley.

GENES/PRIMER: nb2-rdg2a (1500bp-PCR product)

nb2-rdg2a_F2	TCCACTACCCGAAAGGCACTCAGCTAC
nb2-rdg2a_R2	GCAATGCAATGCTCTTACTGACGCAAA

TILLING PCR REACTIONS (TaKaRa ExTaq enzyme): total volume: 25uL

10x ExTaq buffer (TaKaRa)	2.5 µL
dNTP mix (2.5 mM)	2.0 µL
Primer forward (10 µM)	0.3 µL
Primer reverse (10 µM)	0.3 µL
TaKaRa Taq (5U /µl)	0.1 µL
Barley genomic DNA (5 ng/µL)	5.0 µL
H2O (to 25 µL)	14.8 µL

TILLING PCR cycling program for TILLING ("PCRTM70")
95°C for 2 min;
loop 1 for 8 cycles (94°C for 20 s, 73°C for 30 s, reduce temperature 1°C per cycle, ramp to 72°C at 0.5°C/s, 72°C for 1 min);
loop 2 for 45 cycles (94°C for 20 s, 65°C for 30 s, ramp to 72°C at 0.5°C/s, 72°C for 1 min);
72°C for 5 min;
99°C for 10 min;
loop 3 for 70 cycles (70°C for 20 s, reduce temperature 0.3°C per cycle); hold at 8°C.

2. Mix 10µL of PCR product with 10uL weed digestion mix to a volume of 20µL
3. Incubate at 45°C for 15 min
4. Add 2.5µL of 0.5M EDTA (pH 8.0) – to stop reaction
5. Load a 10µL aliquot on an agarose gel

13.4. Example results

Concentrations of protein extracts:

Table 1. Calculations of concentration factors after centrifugation with Amicon Ultra 10K – Starting volume: 500 µL
("Before" centrifugation = considered as 1x concentrated)

	Recovered volume	**Concentration factor (calculated from 500 µL starting volume)**
Weed	~42 µL	**11.9x**
CelI	~33 µL	**15.2x**

Mismatch digestions using celery and weed enzyme extracts:

Table 2. Set-up of mismatch digestions using celery and weed enzyme before and after centrifugation with Amicon Ultra 10K. The enzyme concentration in the extracts were calculated using the calculated concentration factors from Table 1.

	1 - BEFORE	**2 - after**	**3 - after**	**4 – after**
Enzyme	3.5 uL (1x)	0.5 uL	3 uL	6 uL
CelI buffer	1.5 uL	1.5 uL	1.5 uL	1.5 uL
H2O	5 uL	8.0 uL	5.5 uL	2.5 uL
Tot.Volume	10 uL	10 uL	10 uL	10 uL
Celery enzyme concentration in relation to extract before centrifugation (3.5uL – before = 1x)	**1x**	7.6 uL **2.2x**	45.6 uL **13.0x**	91.2 uL **26.1x**
Weed enzyme concentration in relation to extract before centrifugation (3.5uL - before = 1x)	**1x**	5.95 uL **1.7x**	35.7 uL **10.2x**	71.4 uL **20.4x**

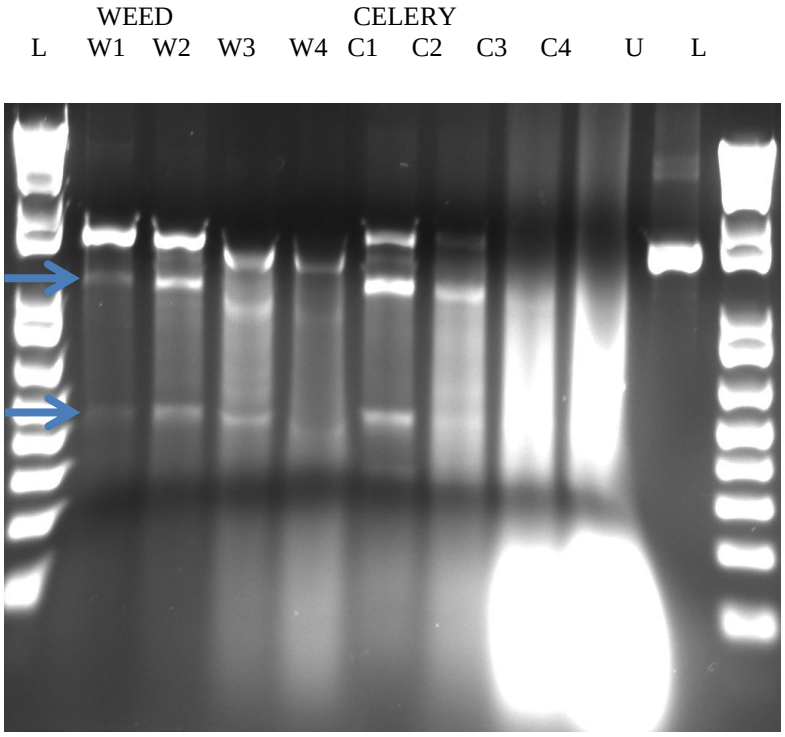


Figure 2. Mismatch cleavage with celery and weed enzyme extracts. TILLING-PCR products of the target gene nb2-rdg2a (1500bp-PCR product) were produced from genomic DNA of barley. The PCR products were digested with weed and celery enzyme extracts before and after concentration by centrifugation with Amicon Ultra 10K. 10 uL of the digested PCR products were separated on a 1.5% agarose gel. Position of SNPs are marked with blue arrows. Concentrations of Weed (W) and Celery (c) extracts are listed above the lanes. A 1kb ladder is loaded on either side of the samples.

13.5. Conclusions

Crude enzyme extracts of weeds show a similar activity to that of celery extract for the cleavge of single nucleotide polymorphisms. The per unit activity, however, was lower than than for CEL I, likely owing to the co-precipitation of other plant proteins in weeds, presumably including a larger amount of RUBISCO. This limitation can be overcome through the use of a simple centrifugation based protein concentration step. 150 mls of weed extract produces enough enzyme for approximately 2000 reactions with this protocol.

14. LOW-VOLUME, NON-TOXIC AND RAPID EXTRACTION OF SINGLE-STRAND-SPECIFIC NUCLEASES FROM CELERY

14.1 Objective

The aim of this protocol is to provide a quick method for crude celery juce (CJE) extraction for 5000 reactions or more that removes the toxicity and use of specialzed equipment and methods (preperatory centrifuge and dialysis), so that it can be performed in a standard molecualar biology laboratory. This enzyme is used for SNP and small indel discovery and genotyping applications such as TILLING and Ecotilling.

14.1. Materials

1. Juicer (*e. g.,* Le Quipe).
2. Celery.
3. 1M Tris-HCl, pH 7.7.
4. Buffer A: 0.1 M Tris-HCl, pH 7.7,
5. Buffer B: 0.1 M Tris-HCl, pH7.7, 0.5 M KCl,
6. Amicon Ultra 0.5ml 10K Centrifugal filters (Millipore Amicon Ref.No. UFC501024 24Pk).

14.2. Methods

14.2.1. CEL I preparation

1. Perform all steps at 4^{o}C when possible. Most steps can be performed at room temperature.
2. Rinse desired amount of celery with water. One bunch (~1 lb) yields approximately enough CEL I for 500,000 standard TILLING reactions. Remove any leaves and cut off tough tissue at base of stalk. For this protocol we aim for the production of about 15mls of juice with 0.5kg of material typically giving 200-400mls.
3. Juice the desired amount of material.
4. Add 1M Tris to a final concentration of Buffer A.
   18 mL *celery juice + 2 mL 1M Tris-HCl buffer (pH=7.7)*
5. Distribute liquid into 10 2.0 ml microcentrifuge tubes.
6. Spin the juice for 20min at 2600 g to pellet debris at 4C if possible.
7. Pour supernatant into a beaker.
8. Bring the supernatant to 25% $(NH_4)_2SO_4$ by adding 144 g per liter of solution. Mix gently at 4^{o}C for 30 min. Using stir plate and magnetic stir bar.

Total volume (from 10 tubes): 18.5 mL – 2.66g $(NH_4)_2SO_4$ added

Figure 1. Protein precipitation with $(NH_4)_2SO_4$

9. Distribute liquid into 10 2.0 ml microcentrifuge tubes. Spin at 15000 g at 4°C for 40 min.

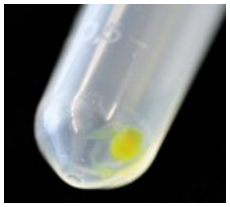


Figure 2. Protein pellet after 25% $(NH_4)_2SO_4$ precipitation (discard)

10. Pour supernatant into clean beaker. Discard pellet.

11. Bring the supernatant from 25% to 80% $(NH_4)_2SO_4$ by adding 390 gram per liter of solution. Mix gently at 4°C for 30 min.
    Total volume (from 10 tubes): 18 mL – 7.02g $(NH_4)_2SO_4$ added

12. Distribute liquid into ~~10~~ 11 2.0 ml microcentrifuge tubes. Spin 15000 x g for 1.5 hr. Save the pellet and discard the supernatant, being careful in decanting the supernatant. The pellet can be stored at –80°C for months.

Figure 3. Protein pellet after 80% $(NH_4)_2SO_4$ precipitation (keep and resuspend)

13. Resuspend the pellets in ~ 1/10 the starting volume with Buffer B, ensuring the pellet is thoroughly resuspended. Target final volume for all 10 pellets is 1.5mls. Add

1.5mls buffer B to tube #1, resuspend by pipetting up and down or vortexing. Then transfter this liquid to tube #2 and repeat, continue until the last tube.

In total it were 11 tubes with pellets: I resuspended 5 pellets in 750uL and 6 pellets in 750uL – then combined to 1.5 mL (total volume of liquid + pellets ~2 mL)

Figure 4. ~2 mL liquid after re-suspension and combination of 11 pellets (derived from 80% $(NH_4)_2SO_4$ precipitation

14. Desalting: Use Amicon ultra filters. Distribute liquid into four filter devices, making sure not to exceede 500ul in any filter. Attach collection tube and spin at 14000g for 30minutes. When complete, remove liquid from collection tube and add 500ul buffer B and repeat. Repeat this step a total of 4 times.

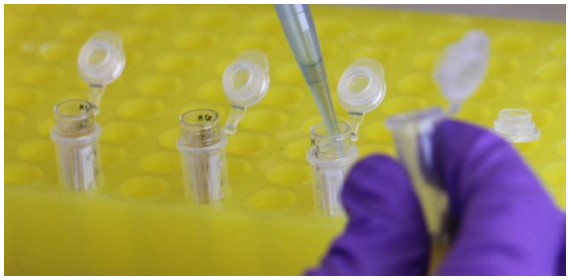


Figure 5. Transfer of liquid after resuspension to 4 Amicon Ultra filter devices

Table 1. Volume of retained liquid in the Amicon Ultra filter devices after the 5 centrifugation steps and a resulting (calculated) desalting factor.

Centrifugation (30 min at 15000 g)	*Starting volume*	*End volume (+ buffer added)*	*Desalting*	*Desalting factor (calculated)*
1	*500*	*100 + 400*	*5x*	*5x*
2	*500*	*~50 + 450*	*10x*	*50x*
3	*500*	*~35 + 470*	*14x*	*700x*
4	*500*	*~30 + 475*	*16.6x*	***11620x***
5	*500*	*For elution*	*no*	-

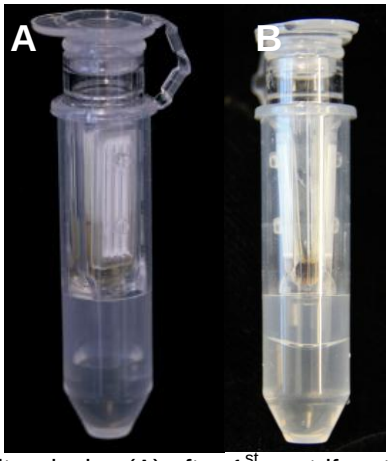


Figure 6. Amicon Ultra filter device (A) after 1st centrifugation (~100uL liquid retained in the filter device and ~400uL flow-through) and (B) after 5th centrifugation (~35uL liquid retained in the filter and ~465uL flow-through)

15. Elute sample: To elute sample, invert the filter and place inverted into a new collection tube. Centrifuge at 1000g for 2min.

Table 2. Volumes of recovered liquid from each Amicon Ultra filter device after inverted centrifugation and (calculated) concentration factor of the enzyme.

Eluate	*Starting vol*	*Elution volume*	
1		*35*	
2		*45*	
3		*40*	
4		*35*	
Total	*2000 uL*	***155 uL***	***12.9 x***

16. Combine all eluates

Final volume of 4 tubes: 155 uL

Figure 7. Recovered eluates after centrifugation of the inverted Amicon filter devices

17. Centrifuge (remove solid material) - 4°C for 30 min at 10.000g

Centrifuged an aliquot of 70uL (other 70 uL frozen without centrifugation)

18. Use supernatant for activity test

14.2.2. Activity tests

For standard TILLING applications, test a range of amounts of CEL I for activity with known mutations/polymorphisms following the high throughput TILLING protocol. Target amount per reaction X = $7.5x10^{-8}$ x total amount juice in µl. For example, the target range for a bunch of celery giving 400mls juice is 400000 x $7.5x10^{-8}$ or 0.03 µl per reaction. To assay activity, perform a standard titration curve with the outermost points flanking the target range on either side by a factor of 100. With excess enzyme, full length PCR product will disappear; as the amount of enzyme falls below the target range, PCR product and background bands will become increasingly dark to the point where the image becomes difficult to interpret.

Synthesis of TILLING-PCR products for mismatch cleavage tests (barley TILLING primer #13)

GENES/PRIMER: Mlo9 (1476 bp-PCR product)

#13	HV_Mlo9-F2	CATTTGTCGCAAAACAGCAAGTTCGAC
	HV_Mlo9-R2	TTGTCTCATCCCTGGCTGAAGGAAAAA

TEMPLATE: 1:1-mixture of Golden Promise and HOR-1606 gDNA – mixture gives mismatch cleavage

TILLING PCR REACTIONS (TaKaRa ExTaq enzyme): total volume: 25uL

10x ExTaq buffer (TaKaRa)	2.5 uL
dNTP mix (2.5 mM)	2.0 uL
Primer forward (10 uM)	0.3 uL
Primer reverse (10 uM)	0.3 uL
TaKaRa Taq (5U /ul)	0.1 uL
Barley genomic DNA (5 ng/uL)	5.0 uL
H2O (to 25 uL)	14.8 uL

TILLING PCR cycling program for TILLING (“PCRTM70”)

95°C for 2 min;
loop 1 for 8 cycles (94°C for 20 s, 73°C for 30 s, reduce temperature 1°C per
cycle, ramp to 72°C at 0.5°C/s, 72°C for 1 min);
loop 2 for 45 cycles (94°C for 20 s, 65°C for
30 s, ramp to 72°C at 0.5°C/s, 72°C for 1 min);
72°C for 5 min;
99°C for 10 min;

loop 3 for 70 cycles (70°C for 20 s, reduce temperature 0.3°C per cycle); hold at 8°C.

- CELI-digestions: mix 10uL of PCR product with 10uL digestion mix to a volume of 20uL (see Table 3 for set-up of digestion mixes)
- Incubate at 45^0C for 15 min
- Add 2.5uL of 0.5M EDTA (pH 8.0) – to stop reaction
- Load a 10uL aliquot on an 1.5% agarose gel

SERIAL DILUTIONS OF CELI enzyme

Table 3. Serial dilutions of isolated CelI enzyme and set-up of Cel digestion mixes.

Dilution factor		**Enzyme (uL)**	**CelI buffer**	**H2O (uL)**
(A) CELI from dialysis	**1x**			
0x		0	1.5 uL	8.5
0.1x		0.35 (1:10)	1.5 uL	8.15
1x	**= 0.35 uL**	**3.5 (1:10)**	1.5 uL	5
5x		1.75	1.5 uL	6.75
10x		3.5	1.5 uL	5
24x		8.5	1.5 uL	0
(B) CELI from Amicon filters*	Dilution factor – **12.9x**			
0x		0	1.5 uL	8.5
0.1x		0.25 (1:100)	1.5 uL	8.25
1x	**= 0.027 uL**	**0.25 (1:10)**	1.5 uL	8.25
5x		1.25 (1:10)	1.5 uL	7.25
10x		2.5 (1: 10)	1.5 uL	6.0
20x		0.5	1.5 uL	8.0
50x		1.25	1.5 uL	7.25
100x		2.5	1.5 uL	6.0
*1x concentration of the CELI purified with Amicon filter devices were calculated using the concentration factor 12.9x				

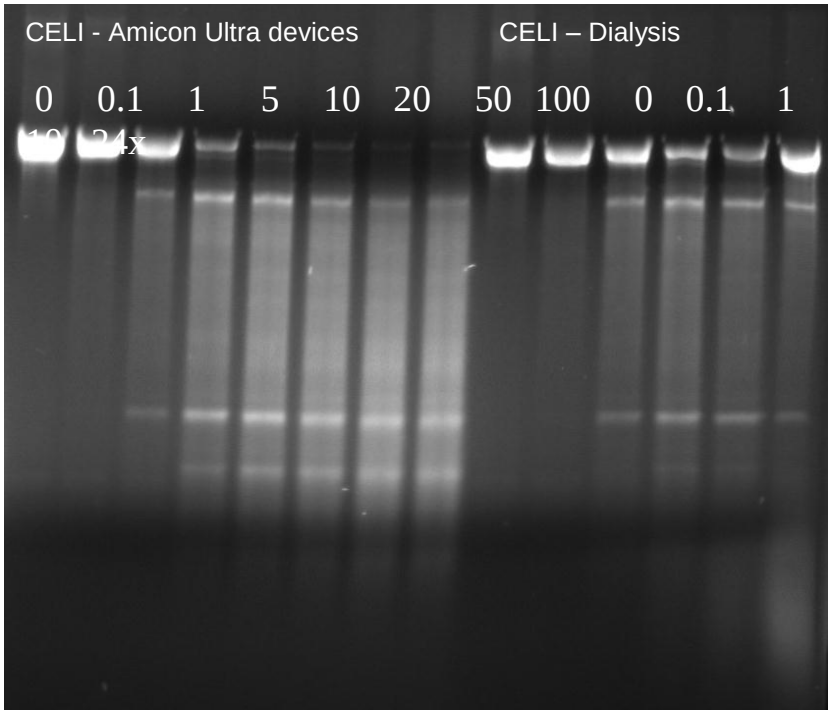


Figure 8. Activity tests of CELI enzyme isolated with Amicon Ultra filter devices (left) and with dialysis method (right). A serial dilution of enzyme activity is shown. There are no cleavage bands present at the control without CELI enzyme (0x) and at the lowest dilution (0.1x) in both extracts. Both enzyme extracts show activity from 1x diluted. However, the background in the CELI extracts purified with Amicon filter devices seems to be stronger than in the activity assays of CELI enzyme purified with dialysis.

14.3. Conclusions

The activity tests showed that mismatch cleavage activity could be detected in celery extracts purified with Amicon Ultra (0.5mL, 10k) centrifugal filter devices (and omitting the dialysis step). The whole isolation procedure could be carried out within 1 day using standard laboratory equipment (i.e. a cooled microcentrifuge). However, a stronger background on the agarose gels (possibly originating from salt remnants retained in the filter devices) is an issue.

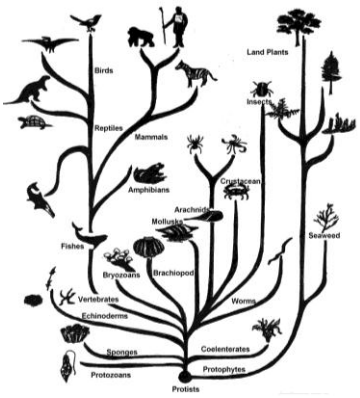
Number of reactions (obtained from 18 mL celery juice and using 4 Amicon Ultra filter devices): we have recovered a total volume of 155 uL from the 4 filter devices. The activity assay shows a clear cleavage pattern with the 5x concentrated enzyme (0.125 uL per reaction). This would allow a total of at least 1240 reactions. However, a lower amount of enzyme (between 1x and 5x) seems to work either and would increase the number of reactions accordingly.

14.4. Contributors

Experimental design, data interpretation: Bernhard Hofinger and Bradley Till
Experimental execution: Bernhard Hofinger
Celery enzyme extraction: Bernhard Hofinger, Owen Huynh, Biguang Huang, Bradley Till
Manuscript preparation and editing: Bernhard Hofinger and Bradley Till
Development of TILLING protocol: Owen Huynh, Bradley Till

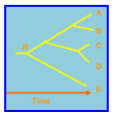
15. MULTIVARIATE ANALYSIS – PHYLOGENETICS AND PRINCIPAL COMPONENT ANALYSIS

15.1. Phylogenetics

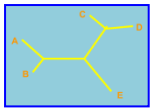


Phylogenetics in the plant kingdom is based on genetic information from accessions. The entities whose affinities are studied are called operational taxonomic units (OTUs, anything from a population to a phylum, including sequence variation and other polymorphisms). Phylogenetics studies the evolutionary relatedness among OTUs using genetic information and is mostly based on genetic distances calculations. The results of these calculations are often synoptically presented as a phylogenetic tree (rooted) or dendrogram (unrooted). There are many methods using different models and assumptions on which the genetic distances calculations are based and ultimately the phylogenetic tree. It is important to understand from the outset what model and *apriori* assumptions to apply in order to be able to infer valuable information from the raw data to be mined.

There are two different tree types that might be constructed, based on two different purposes in analysing the raw data:



Rooted trees serve to unfold an evolutionary path



Un-rooted trees (dendrograms) are used to visualize relationships

A multitude of tree reconstruction algorithms are available. These can be roughly classified into 4 methods:

- Distance Matrix, based on pairwise evolutionary distances (*e.g.* UPGMA, Neighbour Joining)
- Maximum Parsimony, based on the shortest pathway to the present character state
- Maximum Likelihood, based on choosing the tree with the largest ML value of the character state presented
- Invariants, based on functions of characters that have an expected value of 0 in some trees and non-zero expectation in other trees.

15.2. Inferring phylogeny from pairwise distances: construction of a distance tree using clustering with the unweighted pair group method with arithmetic mean (UPGMA).

There are mainly two multivariate methods widely used for pattern analyses of DNA genotypes in biology: principal component analysis (PCA) (Flury 1988) and cluster analysis (Everitt 1992). PCA and cluster analysis seek to uncover hidden or cryptic patterns among objects (*e.g.*, individuals, genetic stocks, or populations) on which two or more independent variables (phenotypic or genotypic characters) have been measured.

• Typical phenotypic variables are morphological traits (*e.g.*, flower petal length and width).

• Typical genotypic variables are DNA marker genotypes or allele sequences. A variety of DNA markers can be employed for genotyping or DNA fingerprinting.

PCA and cluster analysis seek to project multivariate phenotypic or genotypic measurements in lower dimensional spaces so that the underlying patterns or structures can be described and visually displayed. The 'genetic' patterns among a set of OTUs (entities, genetic materials) usually cannot be directly discerned from DNA fingerprints (raw multivariate data); however, patterns among the OTUs can nearly always be 'extracted' by PCA or cluster analyses of pairwise genetic distance matrices.

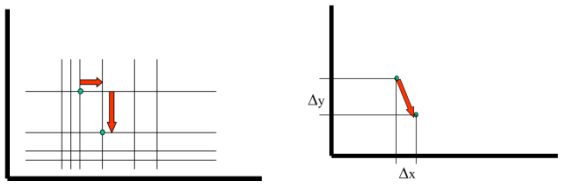
Originally developed for constructing taxonomic phenograms, *i.e.* trees that reflect the phenotypic similarities between OTUs, UPGMA is the simplest method of tree construction, if the rates of evolution are approximately constant among the different lineages. For this purpose the number of observed nucleotide or amino-acid substitutions can be used.

15.3. Distance measures

Distance measures are based on topology paths in n-dimensional space. As an example in a two dimensional space we might consider the following:

Travel in a grid *versus* shortest direct distance

$d = \min(\Sigma x_i + \Sigma y_j)$ $\quad$ $d = SQR(\Delta x^2 + \Delta y^2)$



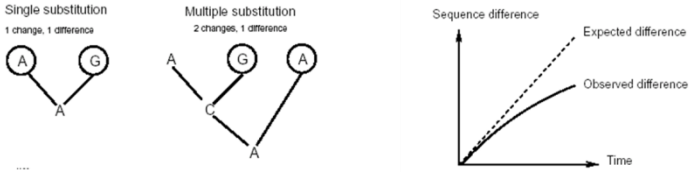
In the context of plant production and protection, the choice of genetic distance estimators depends on what we want to do, what we want to see, what precision of their estimations is needed and the conditions of their applications (in terms of type of markers, genetic structure of the cultivars/accessions/individuals, diversity of reference collections, breeding programmes *etc*). This defines the dimensions and topologies of the space we are exploring and the paths in this space. Let us construct the following set-up to illustrate the

utmost importance of the choice of a genetic distance estimator (*i.e.* it should not be chosen uniquely given the availability of a computer programme): a "naïve" measure of genetic similarity or measure of genetic distance is the Hamming distance where d_0 = proportion of sites at which two sequences differ:

Sorghum TGTATCGCTC…
Sugarcane TGTGTCGCTC…

Sorghum TGTATCGCTC…
Rice AGTCTCGTTC…

Sugarcane TGTGTCGCTC…
Rice AGTCTCGTTC…



The Hamming Distance is a poor measure of the actual number of evolutionary changes, as a site can undergo repeated substitutions. It might be appropriate for short periods and/or parental inferences.

In order to define a genetic distance estimator, we have to assay the genetic similarities of the entities we are studying. Let these entities be dominant markers (present-absent characters): the genetic similarity between the *i*th and *j*th entity is s_{ij}. As such, genetic similarity coefficients are symmetric ($s_{ij} = s_{ji}$), positive and bound by 1 ($0 \leq s_{ij} \leq 1$). Two individuals are completely identical, when $s_{ij} = 1$ and completely different when $s_{ij} = 0$
Genotypic scores and counts for a binary variable (dominant marker):

entity i		**entity j**	**count**			**condition**
present	(1)	present	(1)	*a*	(n_{11})	positive match
present	(1)	absent	(0)	*b*	(n_{10})	mismatch
absent	(0)	present	(1)	c	(n_{01})	mismatch
absent	(0)	absent	(0)	*d*	(n_{00})	negative match

The two most widely used similarity measures for binary data are the simple matching coefficient and Jaccard's coefficient.

- The simple matching coefficient is the ratio of the sum of matches to the sum of matches and mismatches:

$$s_{ij} = \frac{a + d}{a + b + c + d}$$

- Jaccard's coefficient is the ratio of positive matches to the sum of positive matches and mismatches:

$$s_{ij} = \frac{a}{a + b + c}$$

Based on defined genetic similarity coefficients, genetic distance measures can be inferred. The Euclidean genetic distance between the *i*th and *j*th entity is:

dij = *SQR[*2(1 – *sij*)], if the genetic similarity matrix is positive semi-definite (Gower 1971). Both simple matching coefficient and Jaccard's coefficient matrices are positive semi-definite.

In linear algebra, a positive-definite matrix is a Hermitian matrix which in many ways is analogous to a positive real number. The notion is closely related to a positive-definite symmetric bilinear form. In mathematics, a definite bilinear form is a bilinear form B such that B(x, x) has a fixed sign (positive or negative) when x is not 0.

To give a formal definition: let K be one of the fields R (real numbers) or C (complex numbers). Suppose that V is a vector space over K, and $B : V \times V \to K$ is a bilinear form which is Hermitian in the sense that B(x, y) is always the complex conjugate of B(y, x). Then B is called positive definite if $B(x, x) > 0$ for every nonzero x in V. If $B(x, x) \geq 0$ for all x, B is said to be positive semidefinite.

A Hermitian matrix (or self-adjoint matrix) is a square matrix with complex entries which is equal to its own conjugate transpose $a_{i,j} = a^*_{j,i}$ — that is, the element in the *i*th row and *j*th column is equal to the complex conjugate of the element in the *j*th row and *i*th column, for all indices *i* and *j*. Or written with the conjugate transpose: $A = A^\dagger$

For example, $\begin{bmatrix} 3 & 2+i \\ 2-i & 1 \end{bmatrix}$ is a Hermitian matrix. For all non-zero $x \in R^n$ (or, equivalently, all non-zero $x \in C^n$), it is called positive-semi-definite if $x^*Mx \geq 0$.

The three most common distance estimators which are computed throughout the majority of the literature for different purposes are: the Jaccard's distance (J) (1908), the Nei & Li's distance (NL) (1979) and the Sokal & Michener's distance (SM) (1958):

$J_{xy} = 1 - (n_{11} / (n_{11} + n_{10} + n_{01}))$ [1]

$NL_{xy} = 1 - ((2 \times n_{11}) / ((2 \times n_{11}) + n_{10} + n_{01}))$ [2]

$SM_{xy} = 1 - ((n_{11} + n_{00}) / (n_{11} + n_{10} + n_{01} + n_{00}))$ [3]

where n_{11} is the number of bands shared by the individuals (cultivars, clones accessions etc.) x and y tested (i.e. positive matching between pairs), n_{10} is the number of bands present in x and absent in y, n_{01} the number of bands present in y and absent in x, and n_{00} the number of bands absent both in x and y (*i.e.* negative matching). In addition, one may also, using the inverse of the PIC (polymorphism information content of a certain marker), compute a weighted Jaccard's distance (WJ) to take into account the frequency of each marker in the calculation of the distance.

$$PIC = 1 - \sum_{i=1}^{n} P_i^2 - \sum_{i=1}^{n} \sum_{j=i+1}^{n} P_i^2 P_j^2 \quad [4]$$

Pi = frequency of allele i from 1 to n

This formula produces an indicator of how many alleles a certain marker has and how much these alleles divide evenly. For example if a marker has few alleles, or if the marker has many alleles but only one of them is frequent, the PIC will be low. Obviously:

$$1 = \sum_{i=1}^{n} Pi \quad [5]$$

The Nei & Li genetic distance estimator was developed for the analysis of restriction site polymorphisms, and is the estimator proposed by Dice (1945) in the pre-molecular era: $D_{ij} = 2N_{ij}/(N_i + N_j)$, where N_{ij} is the number of restriction sites or restriction fragments shared by *i* and *j* ($= n_{11}$), N_i is the number of restriction fragments in *i* ($n_{11} + n_{10}$), and N_j is the number of restriction fragments in *j* ($n_{11} + n_{01}$). This estimator excludes negative matches.

The simple matching coefficient and Jaccard's coefficient differ in how negative matches (0-0 matches or d counts) are handled. The problem of whether to include or exclude negative matches only arises for present-absent characters (binary or categorical variables), e.g., binary genetic markers with null alleles.

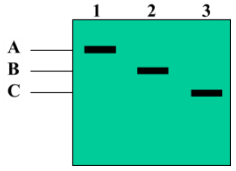
The question as to whether two individuals are similar when they both lack a character does not always have a simple answer. This topic has been hotly debated, particularly in taxonomic circles (Romesburg 1984; Sneath and Sokal 1973). When one allele is absent (null) and the other is present and both alleles are observed among the entities sampled, Dudley (1993) argued that 0-0 matches should be included because the absence of an allele in two entities measures similarity. This may or may not be true. Two individuals, for example, may lack an AFLP band; however, the mutations that abolished the AFLP band in the two individuals could be different (mutation in the restriction sites = elimination of sites, insertion between restriction sites = band too long to amplify, deletion between restriction sites = smaller band appearing but too small to be scored, translocation = reshuffling restriction sites), in which case the two individuals carry different null alleles and the 0-0 score is incorrect. But the probability of these events locus by locus depends on the frequency of these events, and the probability of loss of band due to different mutation events decreases with increasing relatedness. In fact, including 0-0 matches increases homoplasy: loci identical by state but not identical by descent. Thus, when estimated from multiallelic markers, genetic similarities may be upwardly biased by including negative matches, particularly when one or more alleles are rare.

Negative matches should be excluded for multiallelic, co-dominant markers with no null alleles, otherwise, similarities are overestimated. In the following, an example illustrating this will be detailed:

Suppose three lines are genotyped for a locus with three codominant alleles and each line is homozygous for a different allele

Entity	Allele 1	Allele 2	Allele 3	Genotype
1	1	0	0	1
2	0	1	0	2
3	0	0	1	3

(1 = present, 0 = absent)



Now, Gower (1971) proposed a similarity measure for cases where mixed variable types are measured (*e.g.*, mixtures of binary, ordinal, categorical, and continuous variables). This coefficient can be used, for example, to combine dominant (binary) and multiallelic, co-

dominant (categorical) DNA markers or discrete genotypic and continuous phenotypic variables and is one of several similarity measures used in genetic pattern analysis. Gower's coefficient and Jaccard's coefficient are the same when the former is estimated from binary variables and negative matches are excluded. We can use this to illustrate whether negative matches should be included or not.

Gower's coefficient is:

$$s_{ij} = \left(\sum_{k=1}^{m} w_{ijk} \times s_{ijk}\right) / \left(\sum_{k=1}^{m} w_{ijk}\right)$$

where the similarity between the *i*th and *j*th entity measured on the *k*th variable is s_{ij}, the weight for the *k*th variable measured on the *i*th and *j*th entity is w_{ij}, $i = 1, 2, ..., n$, $j = 1, 2, ..., n$, n is the number of entities, $k = 1, 2, ..., m$, and m is the number of variables (DNA fragments or bands). The variable weight is either 0 or 1 and is used to include or exclude negative matches for binary or categorical variables (genetic markers). when k is unknown for one or both entities.

In our example, if we exclude 0-0 matches:

Outcome	Entity *i*	Entity *j*	s_{ijk}	w_{ijk}
if positive match	1	1	1	1
if mismatch *i* - *j*	1	0	0	1
if mismatch *i* - *j*	0	1	0	1
if negative match	0	0	1	0

$s_{12} = ((0 \times 1) + (0 \times 1) + (1 \times 0))/(1 + 1 + 0) = 0/2 = 0$
$s_{13} = ((0 \times 1) + (1 \times 0) + (0 \times 1))/(1 + 0 + 1) = 0/2 = 0$
$s_{23} = ((1 \times 0) + (0 \times 1) + (0 \times 1))/(0 + 1 + 1) = 0/2 = 0$

Now, if we include 0-0 matches:

Outcome	Entity *i*	Entity *j*	s_{ijk}	w_{ijk}
if positive match	1	1	1	1
if mismatch *i* - *j*	1	0	0	1
if mismatch *i* - *j*	0	1	0	1
if negative match	0	0	1	1

$s_{12} = ((0 \times 1) + (0 \times 1) + (1 \times 1))/(1 + 1 + 1) = 1/3$
$s_{13} = ((0 \times 1) + (1 \times 1) + (0 \times 1))/(1 + 1 + 1) = 1/3$
$s_{23} = ((1 \times 1) + (0 \times 1) + (0 \times 1))/(1 + 1 + 1) = 1/3$

The genetic similarities among the lines (considering the one locus only) are 0.00; however, if negative matches are included, then the genetic similarities are 0.33.

Obviously this is a "demonstration by the absurd": we have a population of 3 entities, genotyping is based on 1 co-dominant locus, there are only 3 alleles in our population, allele frequency of all the alleles is identical in our population, and we are sure that there is no null

allele, further all individuals are homozygotes. Obviously genetic similarities should be 0. In our thought experiment, to include 0-0 matches is wrong.

Unfortunately, in "real" life, matters are not so easy. But our example shows, what questions we have to answer before deciding which model to use: heterozygosity, *a priori* knowledge of the population (structure, phylogeny), allelism (number, frequencies, null-alleles), marker system (dominant/co-dominant). Unfortunately, some of these data cannot be assessed. A fruitful approach, in my opinion, is to compare the results of different models and look for consistencies/differences, which contradict our *a priori* expectations and trying to find an explanation to these puzzles.

In some cases, the simple coefficients of correlation between these four genetic distances (J, NL, SM and WJ) may be calculated, *e.g.* to test whether there is an effect due to the choice of the distance. If the correlation is high for the six pairwise comparisons (e.g. over 0.9), then one might not bother about the biology, reproduction system (vegetatively *versus.* sexually propagated, auto/allogamous), ploidy, heterozygosity or population structure. One has not to forget that genetic diversity analysis is not just "number crunching": it is the knowledge of the plant biology and the characteristics of the used marker system(s) which prompts the choice, eventually the construction, of a mathematical model to analyse the data.

For example: the choice of the euclidean distance leading to Jaccard or Dice-indeces is *a priori* a model to consider when using RAPD markers. The Dice index (Jaccard, euclidean distance) is more robust against artefactual bands, but takes into account only common present bands. Now AFLP is more reproducible than RAPD, and absent bands are very significant indeed, and an algorithm such as the "simple matching algorithm", or an algorithm of Sokal and co-workers is more appropriate.

So when confronted with analysing genetic diversity, one should start by acknowledging the biological characteristics of the plant and the general taxonomy (genera, species e.g.) of the individuals/accessions in the study (assess the *a priori* structure of the genetic diversity of a collection of individuals, phenotyping). Then look into the characteristics of the marker system(s) used: dominant *vs.* co-dominant, PIC, reproducibility (confidence in reading the pattern, power of resolution of the analysis system, for example). This will prompt a choice of different mathematical models applicable to the problem, or even more interestingly exclude some choices.

In general: the choice of the Dice-index is at least worth a tentative first order approximation to genetic diversity analyses to sketch a rough outline of genetic diversity of the population studied. To confirm/refine this working draft (compare/oppose the *a priori* structure of genetic diversity to the one obtained using the Dice-index), one might have to use co-dominant markers to assess ploidy, heterozygosity. This might bring new insights furthering data re-analyses using more appropriate algorithms, adapted to the plant biology and/or marker characteristics, to get a better modelisation of diversity.

The sampling distributions of genetic distance estimators are not known; thus, parametric methods for estimating sampling variances and constructing confidence intervals have not

been developed; however, bootstrapping or other resampling methods can be used to estimate sampling variances. Bootstrapping is done by randomly sampling data with replacement to produce individual samples from which the parameters are estimated. Suppose *n* individuals were sampled from a population to estimate allele frequencies. Bootstrapping would be done by drawing *b* bootstrap samples of *n* individuals with replacement and producing *b* allele frequency estimates from which mean allele frequencies and sampling variance are estimated.

When constructing dendrograms bootstrapping generates multiple data sets (usually 100 random resampling iterations with replacement are sufficient, format of seed number being [4n+1]) and adds statistical significance to the branching points in the dendrograms, which are good starting points for discussions in an article. Sometimes PCA (principal component analysis) eigenvector decomposition into major axes for 2D representation of clustering give a better synoptic background to discussions than dendrograms.

15.4. Some reflexions on the comparison between genetic distances.

NL can be easily expressed as an increasing function of J (NL = J / [2 - J]), which means that one is to expect them to be very highly correlated and lead to identical rankings of genetic distances. If this expectation is not met, this is very significant and needs to be investigated

In comparison, a high correlation between J and SM is not obvious. The difference between these distances (formula [1] and [3]) come from negative matches which are taken into account in the denominator of SM distance.

Peltier *et al.* (1994), supported that in the case of intra-specific studies, an allelic relation exists between presence and absence of a band and a negative matching is an indication of similarity and might lead to the same kind of results with SM and J.

In addition, if the weighting of Jaccard (WJ) distance by the inverse of the PIC provides similar relationships between cultivars/accessions/individuals to Jaccard ones, this might be due to the structure of the marker frequency between individuals tested. But WJ leads to take the most different individuals further away from each other, enhancing differences and might clarify

15.5. What genetic distance estimator to choose for essential derivation?

In the framework of plant production and protection, the choice of the genetic distance is crucial for determining the level of relatedness between cultivars/accessions. For the distinctness and without any genetic consideration, J and NL are independent of the samples because only bands present in x and/or in y are considered. For SM, negative matches are counted and if a new cultivar/accession carries a new band absent in the previously registered ones, this becomes a new negative matching for these cultivars and the distance will change. For pragmatic reasons, the stability of genetic distance is a very attractive quality for breeders

because a distance between two cultivars is constant when the number of cultivars in the reference collection increases.

But on the other side, the disadvantage of J results in the difficulty of finding statistical distribution of this distance which is important to calculate a confidence interval. This difficulty comes from the denominator, which is not a constant but a random variable. It is easier to work with euclidian distances like SM. They can be modelled as a binomial variable and their statistical properties are well known (Dillmann *et al*. 1997).

15.6. Genetic distances between populations

Genetic distance measures between populations are a generalization from the distance measures we have seen above.
Nei's genetic distance between the *i*th and *j*th population, using the notation of Weir (1996), is

$$D_{ij} = -\ln(d_{ij}) = -\ln\left[\frac{\sum_l \sum_u p_{lu_i} p_{lu_j}}{\sqrt{\sum_l \sum_u p_{lu_i}^2 \sum_l \sum_u p_{lu_j}^2}}\right],$$

where p_{lui} is the frequency of allele A_u for locus l in the *i*th population and p_{luj} is the frequency of allele A_u for locus l in the *j*th population.
Nei's genetic identity between the *i*th and *j*th population, corrected for sampling bias (Nei 1978), is

$$D_{ij} = -\ln\left[\frac{(2n-1)\sum_l \sum_u p_{lu_i} p_{lu_j}}{\sqrt{\sum_l \left(2n\sum_u p_{lu_i}^2 - 1\right)\sum_l \left(2n\sum_u p_{lu_j}^2 - 1\right)}}\right],$$

where n is the number of individuals sampled within each population.
Hillis (1984) proposed a genetic distance estimator to overcome the problem of Nei's genetic distance estimator producing greatly different estimates when polymorphisms within populations vary. The Hillis genetic distance estimator is

$$D_{ij} = -\ln\left[\sum_l \left(\frac{\sum_u p_{lu_i} p_{lu_j}}{\sqrt{\sum_u p_{lu_i}^2 \sum_u p_{lu_j}^2}}\right) / m\right],$$

where p_{lui} is the frequency of allele A_u for locus l in the ith population, p_{luj} is the frequency of allele A_u for locus l in the jth population, l = 1, 2, ..., m, and m is the number of loci.

Roger's genetic distance (1972) between the ith and jth population is defined by

$$D_{ij} = \frac{1}{m}\sum_{l}\left[\frac{1}{2}\sum_{u}(p_{lu_1} - p_{lu_2})^2\right],$$

where p_{lui} is the frequency of allele A_u for locus l in the ith population, p_{luj} is the frequency of allele A_u for locus l in the jth population, l = 1, 2, ..., m, and m is the number of loci.
The genetic distance estimators proposed by Nei (1972, 1978) and Rogers (1972) are affected by within population heterozygosity (Swofford *et al*. 1996). Cavalli-Sforza and Edwards (1967) proposed an estimator that overcomes this problem. The arc distance estimator of Cavalli-Sforza and Edwards is:

$$D_{ij} = \sqrt{\frac{1}{m}\sum_{l}\left[\left(2\cos^{-1}\sum_{u}\sqrt{p_{lu_i}p_{lu_j}}\right)/\pi\right]^2},$$

where p_{lui} is the frequency of allele A_u for locus l in the ith population, p_{luj} is the frequency of allele A_u for locus l in the jth population, l = 1, 2, ..., m, and m is the number of loci.
Populations are conceptualised as existing as points in an m-dimensional Euclidean space which are specified by m allele frequencies (i.e. m equals the total number of alleles in both populations). The distance is the angle between these points (chord):

$$d_{chord}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} x_{1j}x_{2j}}{\sqrt{\sum_{j=1}^{p} x^2_{1j}\sum_{j=1}^{p} x^2_{2j}}}\right)}$$

where x_i and y_i are the frequencies of the ith allele in populations X and Y
• If no alleles are shared between populations i and j, then D_{ij}=1, "regardless of the variability within either population" (Swofford *et al*. 1996), a property lacking in the estimators of Nei (1972, 1978) and Rogers (1972).
• The angular transformation of allele frequencies seeks to eliminate the adverse effects of different allele frequency ranges.
Nei's genetic distance estimators are based on the following assumptions: Infinite-Alleles Model, all loci have same rate of neutral mutation, mutation-genetic drift equilibrium, stable/constant effective population size (N_e), linear in time
Cavali-Sforza's genetic distance estimator assumes genetic drift only (no mutation), accommodates changes in population size, is linear ib sum of $1/N_e$ over time

15.7. Protocol: tree reconstruction

UPGMA employs a sequential clustering algorithm, in which local topological relationships are identified in order of similarity, and the phylogenetic tree is built in a stepwise manner. We first identify from among all the OTUs the two OTUs that are most similar to each other and then treat these as a new single OTU. Such an OTU is referred to as a composite OTU.

Subsequently from among the new group of OTUs we identify the pair with the highest similarity, and so on, until we are left with only two OTUs.

Definition

Distance d_{AB} between clusters A, B from individual distances d_{ab} :

$$d_{AB} = \frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d_{ab}$$

The distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU and the constituent simple OTUs of the composite OTU. Then a new distance matrix is recalculated using the newly calculated distances and the whole cycle is being repeated.

Algorithm

Initialisation

- Assign each sequence i to its own cluster C_i. Define one leaf for each sequence, and place at height zero.

Iteration

- Determine the two clusters i, j for which d_{ij} is minimal.
- Define a new cluster $C_k = C_i \cup C_j$
- Define a new node k with daughter nodes i and j, and place it at height $d_{ij}/2$.
- Add k to the current clusters and remove i and j.

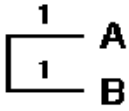
Termination

- When only two clusters i, j remain, place the root at height $d_{ij}/2$.

Following the first clustering A and B are considered as a single composite OTU (A,B) and we now calculate the new distance matrix as follows:
dist(A,B),C = (distAC + distBC) / 2
dist(A,B),D = (distAD + distBD) / 2
dist(A,B),E = (distAE + distBE) / 2
dist(A,B),F = (distAF + distBF) / 2
and so on.

Example
Suppose we have the following distance matrix giving the pair wise evolutionary distances of 6 OTUs:

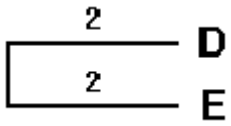
	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8



First cycle
We now cluster the pair of OTUs with the smallest distance, being A and B, that are separated by a distance of 2. The branching point is positioned at a distance of 2 / 2 = 1 substitution. We thus construct a sub-tree as follows:
Following the first clustering A and B are considered as a single composite OTU (A,B) and we now calculate the new distance matrix as follows:
dist(A,B),C = (distAC + distBC) / 2 = 4
dist(A,B),D = (distAD + distBD) / 2 = 6
dist(A,B),E = (distAE + distBE) / 2 = 6
dist(A,B),F = (distAF + distBF) / 2 = 8
In other words the distance between a simple OTU and a composite OTU is the average of the distances between the simple OTU and the constituent simple OTUs of the composite OTU. Then a new distance matrix is recalculated using the newly calculated distances and the whole cycle is being repeated:

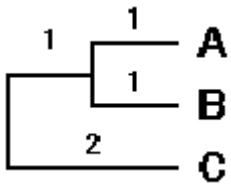
Second cycle

	A,B	C	D	E
C	4			
D	6	6		
E	6	6	**4**	
F	8	8	8	8



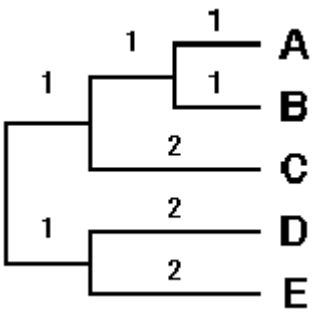
Third cycle

	A,B	C	D,E
C	**4**		
D,E	6	6	
F	8	8	8



Fourth cycle

	AB,C	D,E
D,E	**6**	
F	8	8



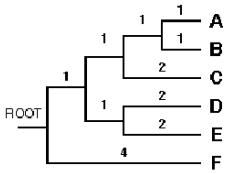
Fifth cycle

The final step consists of clustering the last OTU, F, with the composite OTU.

	ABC,DE
F	**8**

Although this method leads essentially to an unrooted tree, UPGMA assumes equal rates of mutation along all the branches, as the model of evolution used. The theoretical root, therefore, must be equidistant from all OTUs. We can here thus apply the method of mid-point rooting. The root of the entire tree is then positioned at dist (ABCDE),F / 2 = 4.

The final tree as inferred by using the UPGMA:



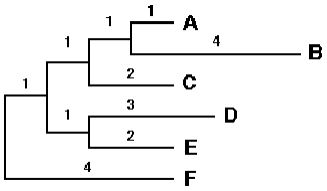
So now we have reconstructed the phylogenetic tree using the UPGMA method. However, there are some pitfalls:

- UPGMA clustering is *very sensitive* to unequal evolutionary rates. This means that when one of the OTUs has incorporated more mutations over time than the other OTU, one may end up with a tree that has the wrong topology.
- Clustering works only if the data are *ultrametric*
- Ultrametric distances are defined by the satisfaction of the '*three-point condition*'.

What is the three-point condition?

For any three taxa: dist AC ≤ max (distAB, distBC) or in words: the two greatest distances are equal, or UPGMA assumes that the evolutionary rate is the same for all branches

If the assumption of rate constancy among lineages does not hold UPGMA may give an erroneous topology. This is illustrated in the following example; suppose that you have the following relationship:



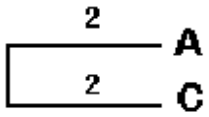
Since the divergence of A and B, B has accumulated mutations at a much higher rate than A. The Three-point criterion is violated! e.g. distBD ≤ max (distBA,distAD) or,

10 ≤ max (5,7) = False

The reconstruction of the evolutionary history uses the following distance matrix:

	A	B	C	D	E
B	5				
C	**4**	7			
D	7	10	7		

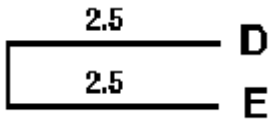
E	6	9	6	5	
F	8	11	8	9	8



We now cluster the pair of OTUs with the smallest distance, being A and C, that are separated a distance of 4. The branching point is positioned at a distance of 4 / 2 = 2 substitutions. We thus construct a sub-tree as follows:

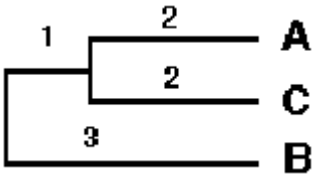
Second cycle

	A,C	B	D	E
B	4			
D	7	10		
E	6	9	5	
F	8	11	8	9



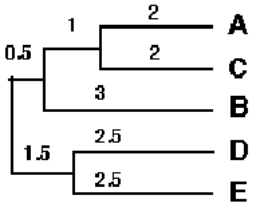
Third cycle

	A,C	B	D,E
B	6		
D,E	6.5	9.5	
F	8	11	8.5



Fourth cycle

	AC,B	D,E
D,E	8	
F	9.5	9.5

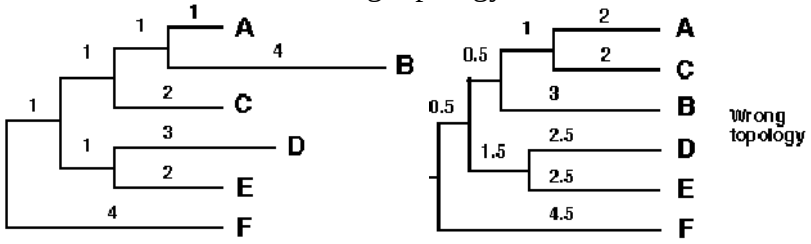


Fifth cycle

The final step consists of clustering the last OTU, F, with the composite OTU, ABCDE.

	ABC,DE
F	9

When the original, correct, tree and the final tree are compared it is obvious that we end up with a tree that has the wrong topology.

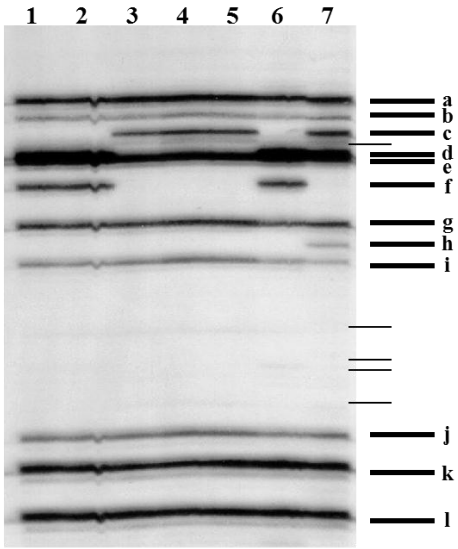


***Conclusion*: The unequal rates of mutation have led to a completely different tree topology.**

15.8. UPGMA exercise

UPGMA exercise

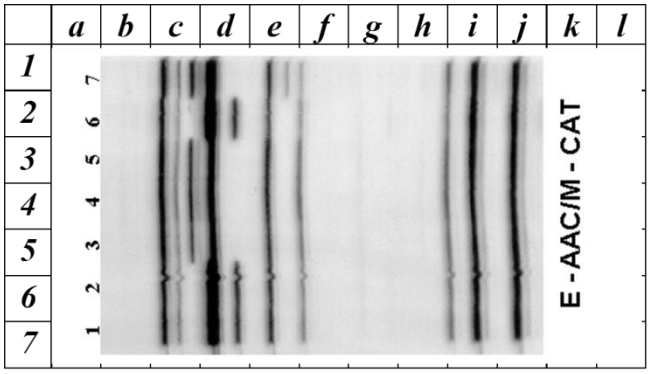
AFLP results



E - AAC/M - CAT

Lane identifications:

- 1= reference 1
- 2= line 01
- 3= line 97
- 4= line 02
- 5= line 09
- 6= line 95
- 7= reference 2



Accessions 2 to 6 were obtained by mutation induction from supposedly accession 1.
Accession 7 is a control.
Verify whether accessions 2 to 6 have been derived from accession 1.

$s_{ij} = (n_{11} + n_{00})/ (n_{11} + n_{00} + n_{10} + n_{01})$

$d_{ij} = sqrt(2*(1- s_{ij}))$

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
1	1	1	0	1	1	1	1	0	1	1	1	1
2	1	1	0	1	1	1	1	0	1	1	1	1
3	1	1	1	0	1	0	1	0	1	1	1	1
4	1	1	1	0	1	0	1	0	1	1	1	1
5	1	1	1	0	1	0	1	0	1	1	1	1
6	1	1	0	1	1	1	1	0	1	1	1	1
7	1	1	1	1	1	0	1	1	1	1	1	1

1:2	n_{11} = n_{00} = n_{10} = n_{01} =											
1:3	n_{11} = n_{00} = n_{10} = n_{01} =	2:3	n_{11} = n_{00} = n_{10} = n_{01} =									
1:4	n_{11} = n_{00} = n_{10} = n_{01} =	2:4	n_{11} = n_{00} = n_{10} = n_{01} =	3:4	n_{11} = n_{00} = n_{10} = n_{01} =							
1:5	n_{11} = n_{00} = n_{10} = n_{01} =	2:5	n_{11} = n_{00} = n_{10} = n_{01} =	3:5	n_{11} = n_{00} = n_{10} = n_{01} =	4:5	n_{11} = n_{00} = n_{10} = n_{01} =					
1:6	n_{11} = n_{00} = n_{10} = n_{01} =	2:6	n_{11} = n_{00} = n_{10} = n_{01} =	3:6	n_{11} = n_{00} = n_{10} = n_{01} =	4:6	n_{11} = n_{00} = n_{10} = n_{01} =	5:6	n_{11} = n_{00} = n_{10} = n_{01} =			
1:7	n_{11} = n_{00} = n_{10} = n_{01} =	2:7	n_{11} = n_{00} = n_{10} = n_{01} =	3:7	n_{11} = n_{00} = n_{10} = n_{01} =	4:7	n_{11} = n_{00} = n_{10} = n_{01} =	5:7	n_{11} = n_{00} = n_{10} = n_{01} =	6:7	n_{11} = n_{00} = n_{10} = n_{01} =	

$$m = \tfrac{1}{2} [(N-1)^2 + (N-1)]$$

The choice of s_{ij} and d_{ij} is given by the problem, (verify relation to parent, AFLP)

Possible simplification based on identity of rows 1, 2 & 6 and rows 3, 4 & 5

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
1	1	1	0	1	1	1	1	0	1	1	1	1
2	1	1	0	1	1	1	1	0	1	1	1	1
3	1	1	1	0	1	0	1	0	1	1	1	1
4	1	1	1	0	1	0	1	0	1	1	1	1
5	1	1	1	0	1	0	1	0	1	1	1	1
6	1	1	0	1	1	1	1	0	1	1	1	1
7	1	1	1	1	1	0	1	1	1	1	1	1

	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*
1,2,6	1	1	0	1	1	1	1	0	1	1	1	1
3,4,5	1	1	1	0	1	0	1	0	1	1	1	1
7	1	1	1	1	1	0	1	1	1	1	1	1

$s_{ij} = (n_{11} + n_{00})/ (n_{11} + n_{00} + n_{10} + n_{01})$

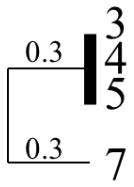
$d_{ij} =$ sqrt(2*(1- s_{ij}))

1,2,6 : 3,4,5	$n_{11} = 8$, $n_{00} = 1$, $n_{10} = 2$, $n_{01} = 1$	$s_{ij} =$ 9/12, $d_{ij} =$ sqrt(6/12) = 0.7
1,2,6 : 7	$n_{11} = 9$, $n_{00} = 0$, $n_{10} = 1$, $n_{01} = 2$	$s_{ij} =$ 9/12, $d_{ij} =$ sqrt(6/12) = 0.7
3,4,5 : 7	$n_{11} = 9$, $n_{00} = 1$, $n_{10} = 0$, $n_{01} = 2$	$s_{ij} =$ 10/12, $d_{ij} =$ sqrt(4/12) = 0.6

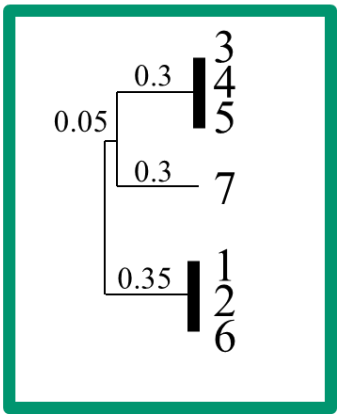
	1,2,6	*3,4,5*	*7*
1,2,6	0		
3,4,5	0.7	0	
7	0.7	0.6	0

three-point condition:
dist(1,2,6:7) <= Max[dist (1,2,6:3,4,5),dist(3,4,5:7)]
0.7<= Max(0.7,0.6) !

	1,2,6	3,4,5	7
1,2,6	0		
3,4,5	0.7	0	
7	0.7	0.6	0



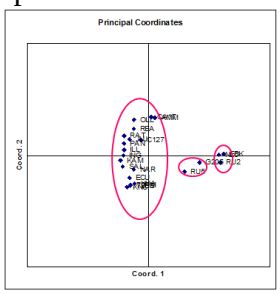
	1,2,6	3,4,5/7
1,2,6	0	
3,4,5/7	0.7	0

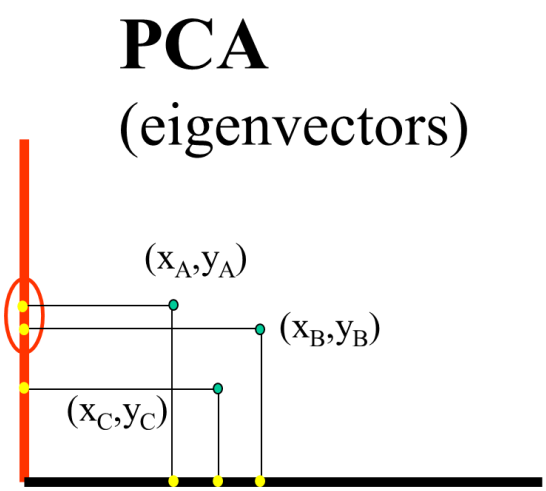


Conclusion:
Mutants 3, 4 & 5 are more related to the control 7 than to the putative parent 1. Possible explanations:
Mislabelling of part of the M_0 and/or M_1
Outcrossing during M_1 selfing

15.9. Principal Component Analysis (PCA)

If a multivariate dataset is represented as a set of coordinates in an n-dimensional data space (1 axis per variable), PCA can reduce the dimensionality of the transformed data and supply a lower-dimensional projection when viewed from its most informative viewpoint, using only the first few principal components. For a seemingly random distribution of data points in the n-dimensional results space, PCA starts with finding the analytical plane by slicing the results space into lower dimensional representations of uncorrelated parameters (eigenvectors).





•finding the analytical plane

In mathematical terms, PCA is a procedure to transform a set of potentially correlated observations into a set of uncorrelated data points: principal components (in number less than or equal to the original variables). This orthogonal transformation is defined in such a way that the first principal component accounts for as much of the variability in the data as possible (maximum variance), and each succeeding component in turn has the highest variance possible under the constraint that it is uncorrelated with (orthogonal to) the preceding components. Principal components are guaranteed to be independent only if the data set is jointly normally distributed.

PCA is the simplest of the true eigenvector-based multivariate analyses. It might be visualized as uncovering the internal structure of the data in a way which best explains their variance. Sensitive to the relative scaling of the original variables, it can be done by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centring the data for each attribute. The results of a PCA are usually discussed in terms of component scores (the transformed variable values corresponding to a particular data point) and loadings (the weight by which each standardized original variable is to be multiplied to get the component score). PCA is closely related to factor analysis; and some statistical packages deliberately merge the two techniques. True factor analysis makes

different assumptions about the underlying structure and solves eigenvectors of a slightly different matrix.

In linear algebra, an orthogonal matrix, is a square matrix with real entries whose columns and rows are orthogonal unit vectors. This means, that a matrix Q is orthogonal if its transpose is equal to its inverse: $Q^T = Q^{-1}$, and thus it follows that $Q^TQ = QQ^T = I$ (I being the identity matrix). An orthogonal matrix Q is thus square, invertible, unitary ($Q^{-1} = Q^*$), and normal ($Q^*Q = QQ^*$). As a linear transformation, an orthogonal matrix preserves the dot product of vectors, and therefore acts as an isometry of Euclidean space, such as a rotation or reflection, thus, it is a unitary transformation.

The eigenvectors of a square matrix are the non-zero vectors that, after being multiplied by the matrix, remain parallel to the original vector. For each eigenvector, the corresponding eigenvalue is the factor by which the eigenvector is scaled when multiplied by the matrix. The prefix eigen- is adopted from the German word "eigen" for "own" in the sense of a characteristic description. In mathematical terms: if A is a square matrix, a non-zero vector v is an eigenvector of A if there is a scalar λ (lambda) such that $Av = \lambda v$

The scalar λ (lambda) is said to be the eigenvalue of A corresponding to v. An eigenspace of A is the set of all eigenvectors with the same eigenvalue together with the zero vector, which however, is not an eigenvector.

15.9.1. Considerations and references

Planning experiments and analyses

Which entities should be sampled?

There are no formal statistical rules for deciding this, so empirical testing is needed. When selecting among a large number of potential entities (*e.g.*, germplasm accessions) or when resources are limiting (which they nearly always are), geographical or ancestral origin, morphological phenotypes, or other phenotypic or historical criteria can often be used to select accessions to represent a gene pool or a specific subset of a gene pool. The genetic material chosen for study depend on economic resources, the nature, scale, scope, and goals of the study, and *a priori* knowledge of genetic relationships. Closely related genetic materials, for example, need not be sampled unless there is a compelling biological or economic reason to do so. The 'ideal' sample of genetic material for studying a particular question is profoundly affected by the nature and genetic origin (if known) of the genetic material.

The goal of a DNA fingerprinting study might be to classify every entity belonging to a particular biological or economic class of entities, *e.g.*, a seed company might fingerprint and classify every inbred line and hybrid they own and every hybrid sold by their competitors for the purpose of protecting intellectual property. Many crop plant gene pools are comprised of hundreds or even thousands of germplasm accessions. Depending on the mating biology and breeding systems of the species, accessions could be comprised of outcrossing wild populations (*e.g.*, genetically heterogeneous, segregating populations), mixtures of inbred genotypes, or inbred lines. How genetically heterogeneous accessions are sampled depends on the goal of the study and economic resources.

Another goal of a DNA fingerprinting study might be to assess the minimum set of accessions that comprise an ideal or so-called core set. The purpose of a core set, in theory, is to produce maximum information from a minimum sample of genetic materials. The practical aims might be to eliminate redundant accessions and streamline the maintenance of genetic diversity in a seed or gene bank.

Similar concepts can be applied to surveys of genetic diversity, *e.g.*, the 'optimum' set of genetic materials for assessing the utility of a sample of genetic markers or, more broadly, for classifying new genetic materials or genetic materials of unknown ancestry or origin.

What is the best sampling strategy?

The mating biology and breeding system of the species dictate the sampling strategy. The gene pools of many plant species, e.g., maize (*Zea mays* L.) and sunflower (*Helianthus annuus* L.), are comprised of partially or 'fully' inbred genetic stocks, in addition to heterogeneous, segregating populations (natural or experimental). The gene pools of humans, most animal species, and many plant species, more or less domesticated and/or wild types, are comprised of heterogeneous, segregating populations.

The optimum genetic and statistical sampling strategies may be difficult to specify, are nearly always constrained by economic factors, and depend on the nature of the statistical analysis and scope of inference. When analyses are performed on segregating populations, a sufficient number of individuals must be sampled within each population to accurately estimate gene

and genotype frequencies. Weir (1996) proposed sampling over loci for random model analyses and over individuals for fixed model analyses. The line between fixed and random is often blurred. Basically, if the scope of inference is across a species or across other strata where broad inferences are to be made, then random models are used. If the scope of inference is a fixed set of populations or inbred lines, then fixed models are used.
If the goal of the study is to survey allelic diversity among a sample of populations (chosen for some biological or commercial reason), then extensive within-population sampling may not be necessary.
If the goal is to accurately describe genetic patterns among populations, measure linkage disequilibrium or gene flow, or protect intellectual property (e.g., an open-pollinated or synthetic cultivars in crop plants), then individuals within populations must be sampled to accurately estimate gene and genotype frequencies and perhaps to find rare alleles and genotypes.

What types of variables should be measured?

Although we are primarily concentrating on the analysis of genotypic measurements (*e.g.*, DNA marker genotypes), phenotypic measurements should not be overlooked and can be combined with genotypic measurements in analyses of genetic patterns. Special similarity measures can be used to combine phenotypic and genotypic measurements or a 'conceptual synthesis' of patterns can be produced from separate analyses performed on phenotypic and genotypic variables. The choice of variables is usually more complicated for phenotypic than genotypic variables, because the former are heterogeneous, whereas the latter are homogeneous (when a single marker system is employed) in the conceptual sense, however, the information supplied by individual genetic markers can vary. If DNA fingerprints are to be produced, then the types of variables measured are dictated (i) by the types of markers developed for the species, (ii) whether the DNA markers are dominant or co-dominant, (iii) by the homology of DNA fragments across individuals or populations, (iv) by economic factors, (v) by the reproducibility and robustness of the DNA marker system (genotyping errors). The ideal genetic marker is highly polymorphic, co-dominant, locus-specific, robust, and highly reproducible.

How many variables should be measured?

There are no formal statistical rules for deciding how many genetic markers are needed to accurately classify accessions, describe genetic patterns, or accurately estimate genetic distances and phenograms.
• Smith *et al*. (1991) used 200 RFLP markers dispersed across the maize genome to fingerprint 11 inbred lines (the genetic distance matrix was comprised of 55 elements). They estimated distance matrices by sampling 5 to 200 RFLP markers in increments of five (*e.g.*, five, 10, 15, ..., 200). They concluded that accuracy was sufficient with 100 or more markers.
• Bernardo (1993) concluded that 250 or more marker loci were needed to produce precise estimates of coefficients of co-ancestry.
The number of genetic markers used in an analysis may be dictated by non-statistical factors. The outcome of the analysis might be one of the criteria used to select genetic markers for future analyses.
Ideally, genetic markers for protecting intellectual property and classifying unknown genetic materials should be highly polymorphic and dispersed across the genome.

Should analyses be performed on raw multivariate data or genetic similarities?

• Typically, multivariate analyses of DNA genotypes (fingerprints) are performed on genetic similarity or distance matrices among entities rather than on raw multivariate data matrices.

• PCA of raw DNA genotypes, although not widely done, can be used to assess the importance of individual genetic markers by comparing principal component coefficients, i.e., individual elements of characteristic vectors (eigenvectors).

15.10. References

(1) **[http://www.icp.ucl.ac.be/~opperd/private/upgma.html]**

Bernardo R. 1993. Estimation of coefficient of coancestry using molecular markers in maize. Theor. Appl. Genet. 85: 1055-1062.

Cavalli-Sforza L.L. and Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 19: 233-257.

Dice L.R. 1945. Measures of the amount of ecological association between species. Ecology 26: 297-302.

Dillmann C., Charcosset A., Goffinet B., Smith J.S.C. and Dattée Y. 1997. Best linear estimator of the molecular genetic distance between inbred lines. In: Krajewski P, Kaczmarek Z (eds) Advances in biometrical genetics. Proceedings of the tenth meeting of the EUCARPIA section biometrics in plant breeding, 14-16 may 1997, Poznan, pp 105-110

Dudley J. W. 1993. Molecular markers in plant improvement: Manipulation of genes affecting qualitative traits. Crop Science (33):660-668 & **Munn R. and Dudley J.** 1995. A PC computer program to generate a dissimilarity matrix for cluster analysis. Crop Sci. 35:925-927.

Everitt B.S. 1992. Cluster analysis. Oxford Univ. Press, New York.

Excoffier L., Smouse P.E. and Quattro J.M. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479-491

Flury B. 1988. Common principal components and related multivariate methods. Wiley, New York.

Gower J.C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27: 857-872.

Hamming R.W. 1950. Error detecting and error correcting codes. Bell System Technical Journal 29 (2): 147–160

Hillis D.M. 1984. Misuse and modification of Nei's genetic distance. Syst. Zool. 33: 238-240.

Hillis D.M., Moritz C., and Mable B.K. 1996. Molecular systematics. Sinauer, Sunderland, Massachusetts.

Jaccard P. 1908. Nouvelles recherches sur la distribution florale. Bull Soc Vaud Sci Nat 44: 223-270

Nei M. 1972. Genetic distance between populations. Am. Nat. 106: 283-292.

Nei M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89: 583-590.

Nei M. and Li W.-H. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. 76: 5269-5273.

Peltier D., Chacon H., Tersac M., Caraux G., Dulieu H. and Bervillé A. 1995. Utilisation des RAPD pour la construction de phénogrammes et de phylogrammes chez Petunia. In: Techniques et utilisations des marqueurs moléculaires. Coll Les colloques INRA

Rogers J.S. 1972. Measures of genetic similarity and genetic distance. Univ. Texas Publ. 7213: 145-153.

Romesburg H.Ch. 1990. Cluster Analysis for Researchers. Florida, Krieger Publishing Co. (original edition 1984).

Smith O.S., Smith J.S.C., Bowen S.L. and Tenborg R.A. 1991. Numbers of RFLP probes necessary to show associations between lines. Maize Genet. Newsltr. 65: 66.

Sneath P.H.A. and Sokal R.P. 1973. Numerical taxonomy. San Francisco, Freeman

Sokal R.P. and Michener C.D. 1958. A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 38: 1409-1438

Swofford D.L., Olsen G.J., Waddell P.J. and Hillis D.M. 1996. Phylogenetic inference. pp. 407-514. Hillis, D.M., C. Moritz, and B.K. Mable (ed.). Molecular systematics. Sinauer, Sunderland, Massachusetts.

Weir B.S. 1996. Genetic data analysis. Sinauer, Sunderland, Massachusetts.

16. POPULATION GENETICS

Population genetics is that branch of genetics that attempts to describe how the frequency of the alleles (of genes) changes over time. To study frequency changes, populations rather than individuals are analysed. The scope of this module however is not to provide an in-depth resource on this branch of science, rather it is aimed at guiding the researcher in a stepwise format through the collection (including coding), analyses and arriving at valid inferences on data for allelic frequencies of molecular markers.

The data coding schemes begin with a random example of a dominant marker gel data. Whether the bands come from RAPD's, ISSR, and AFLP's or similar, does not affect the way data is coded, and more importantly, how it is analysed. What matters, is whether or not *we observe* a given band.

Next, co-dominant markers are dealt with as they are close to the notion of a diploid species where each individual carries *n maternally* and *n paternally* inherited gametes for a total ploidy of 2*n*. Of course, codominant data can be obtained in tetraploid or hexaploid individuals also, as will be demonstrated. The exercises will start with microsatellite data from a population sample. It is important to note however that all these coding systems can be used also for allozyme data. Different coding schemes will be analysed, some 'tricks' with using spread sheets and highlights on what can, and what cannot be done with each coding system will also be shown.

After reviewing how data can be coded, the next step will involve going through the basic concepts of population diversity, population structure, and population divergence. This last part of this module is the basis of phylogenetic studies, although for this manual, only phenetic analyses will be shown.

To conclude this brief introduction to population genetics, two non-exhaustive lists of references and of web-resources of relevance to the study of the subject are provided. Finally, a list of key concepts and equations are provided to complete the definitions given in the text.

16.1. Reading and coding genetic data

16.1.1. Presence/absence coding of dominant data

The most commonly used way for coding genotypes or genetic marker data is by doing a matrix of presence/absence of bands, usually with 1's and 0's. This type of markers is easy to read, provided the number of bands is reasonable and clear. Band intensity, is an issue, and interpretations may change from person to person.

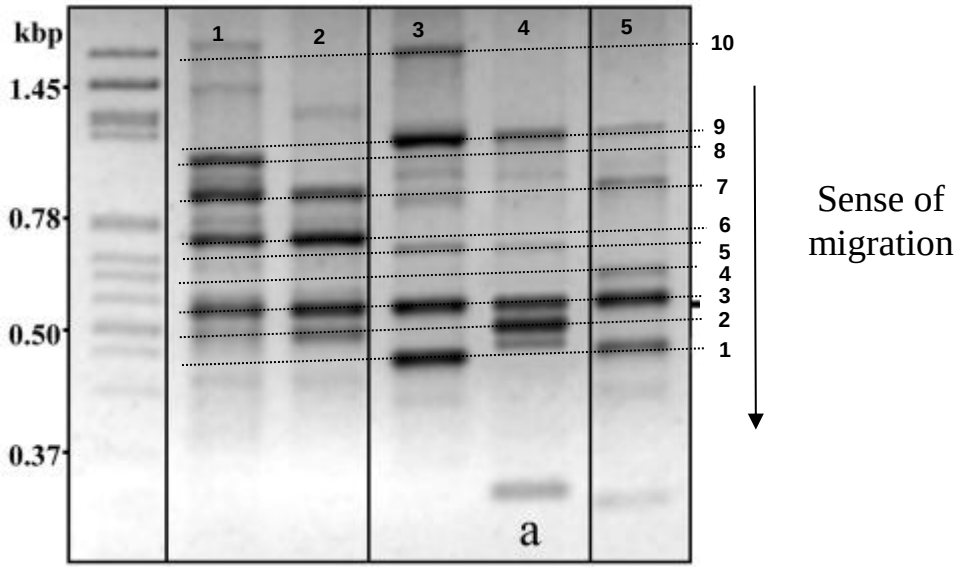


Figure 16-1. Typical dominant data gel, consisting of 5 lanes, and at least 10 well identifiable bands. Bands are scored 1, if present, zero, otherwise. Table 1 shows one reading of this gel into a spread sheet program (interpretation may vary from person to person, or from day to day).

Table 16.1–1. Basic transcription of a dominant marker gel into a spread sheet. Data are organized by columns (fields: id, b1, b10) and individuals are rows (records).

	A	B	C	D	E	F	G	H	I	J	K
1	id	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10
2	1	0	0	1	0	0	1	1	1	0	0
3	2	0	1	1	0	0	1	1	0	0	0
4	3	1	0	1	0	1	1	0	0	1	1
5	4	0	1	1	0	0	0	0	0	1	0
6	5	1	0	1	1	0	0	1	0	1	0
7											

As will be seen later, this coding is not complete for analysis with corresponding software, but is a good starting point. Score bands are highlighted grey for clarity purposes.

16.1.2. Allele size coding for microsatellites

Figure 13.1 shows a typical microsatellite data with 7 alleles in 9 individuals (the number of alleles may change according to the person that reads the gel!). This marker is codominant, because we can see that individuals can bear two alleles at the same time. In principle, each product is originated in the two homologous parts for that particular locus, and if the two alleles are the same, a darker, single band should be seen. Figure 13.2 and Table 13.2 show a first interpretation of this gel in a codominant fashion, upon which inbreeding little *f* or f_{is} can be computed as well as other statistics (see chapters 2 and so forth).

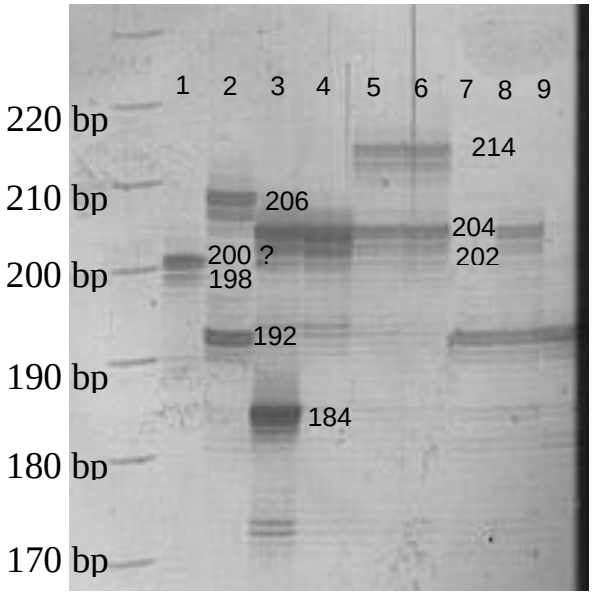


Figure 16-2. Test gel of *Quercus humboldtii* (Andean oak, Colombia) showing 9 individuals (Fernandez, unpublished data). This gel presents many of the typical features of microsatellites: many alleles, stuttering bands, more than two "main" bands, and ambiguity of allele size. A sequencer will also give you results of the type **202.14** bp that the researcher needs to round. Rounding is necessary at this stage or at later steps as most programmes only accept integer numbers.

Table 16.1–2. Same data from example gel using a regular spread sheet programme. Note individuals appear in *rows* (records), and particular data (fields) are in *columns*. Note that individuals 7 and 9 are coded as homozygotes and not as one allele with missing data. Some programs deal with "null" alleles, i.e., false homozygotes due to PCR problems, and in that case, the notation would indicate one un-observed allele.

	1	2	3	4
1	population	individual	locus_1_allele1	locus_1_allele2
2	1	1	198	200
3	1	2	192	206
4	1	3	184	204
5	1	4	202	202
6	1	5	204	214
7	1	6	204	214
8	1	7	192	192
9	1	8	192	204
10	1	9	192	192
11	2	1		
12	2	2		
13	2	3		
14				
15				

16.1.3. Categorical coding

A second interpretation of this gel, would be simply naming the alleles with letters or numbers (preferred coding) from 1 to 8; this is what is usually called "categorical" or "allelic states" coding of alleles that in this case disregards the size information present in the microsatellites (bp's). We will see that the size information is important for genetic distances such as Delta μ^2 and others, but that allelic state is sufficient for genetic distances such as Nei's standard genetic distance, widely used for allozyme data. Figure 13.3 and Table 13.3 shows the coding in "categorical" or "allelic states" for the same gel.

16.1.4. Presence/absence coding of co-dominant data

Yes, you are reading right. A third coding scheme is the popular one that uses 0's (zeros) and 1's (ones), usually called "presence/absence" coding that we just saw for Dominant data in the first section (13.1.1). Often times, we are not interested in evolutionary models and/or samples do not come from random samples from natural populations. We may have accessions coming from different countries or regions within countries collected simply because they present an interesting trait: nice fruits, long spikes, little cyanide, etc. This coding is required for traditional statistics such as Principal Components Analysis (PCA) and related multivariate techniques, with the advantage that genetic data can be combined with morphological data for grouping purposes. Table 4 shows the presence/absence coding for the same example gel.

Important Note: You may notice that this coding is not exclusively for diploids. In fact, tetraploids or hexaploids can be handled this way. Simply, there can be more than two bands per individual, and the notion of heterozygotes diffuses and becomes secondary.

It is clear that for allozyme data, or morphological data known to be co-dominant (white, lilac and purple flowers in *Lynanathus*, for example), "presence/absence" are perfectly applicable. At this point, we would lose the diploid information so estimation of inbreeding (the parameter f_{is} that measures the probability that two alleles within an individual are the same) cannot be computed. This coding, however, is highly popular for analysing accessions because if you will, it is "model" free, and as seen from Table 4, we can include in the same database different kinds of data, and potentially in the same analysis (fruit data color could be changed to 1, 2 and 3 *etc.* to run all in the same analysis, but all depends on the programme used).

Table 16.1–3. Example of Co-dominant data coded as presence absence of bands. First, the total number of alleles is counted, and the corresponding number of bands is defined, being 8 in our case. Note that for homozygote individuals (we are dealing with diploid data) there is controversy about the scoring. In the example below the individuals 7 and 9 were coded as 1 for allele 1, but some people think we should give them twice as much weight (i.e., two copies are there!) so the genotype should be "2" instead of "1". This is no longer "presence/absence" strictly, but results change little in practice.

	1	2	3	4	5	6	7	8	9	10	11
1	population	individual	band_01	band_02	band_03	band_04	band_05	band_06	band_07	band_08	fruit_color
2	1	1	0	0	1	1	0	0	0	0	red
3	1	2	0	1	0	0	0	0	1	0	red
4	1	3	1	0	0	0	0	1	0	0	crimson
5	1	4	0	0	0	0	1	1	0	0	green
6	1	5	0	0	0	0	0	1	0	1	red
7	1	6	0	0	0	0	0	1	0	1	red
8	1	7	0	1	0	0	0	0	0	0	crimson
9	1	8	0	1	0	0	0	0	0	1	red
10	1	9	0	1	0	0	0	0	0	0	red
11	2	1									
12	2	2									
13	2	3									
14											

16.1.5. Formatting dominant data as co-dominant

As strange as it may sound, we can code dominant data as codominant for using a codominant-based data analysis software. Some functions may not work, and the measure of inbreeding will be totally false, but genetic distances, using shared allele distances can be computed. In this case, we would code as follows:

- 22 for the presence of a band
- 11 for the absence of a band
- (-99 if missing data are allowed… not easy to know for dominant markers!)

The file should look similar to that in Table 3 (categorical or allelic state coding) before we transform it in its final form, as shown in Table 5. We can no longer use zeros, because in this context, zeros are usually reserved for missing data!

Table 16.1–4. Dominant data (Figure 13.1) coded as codominant *i.e.*, 2-alleles per band.

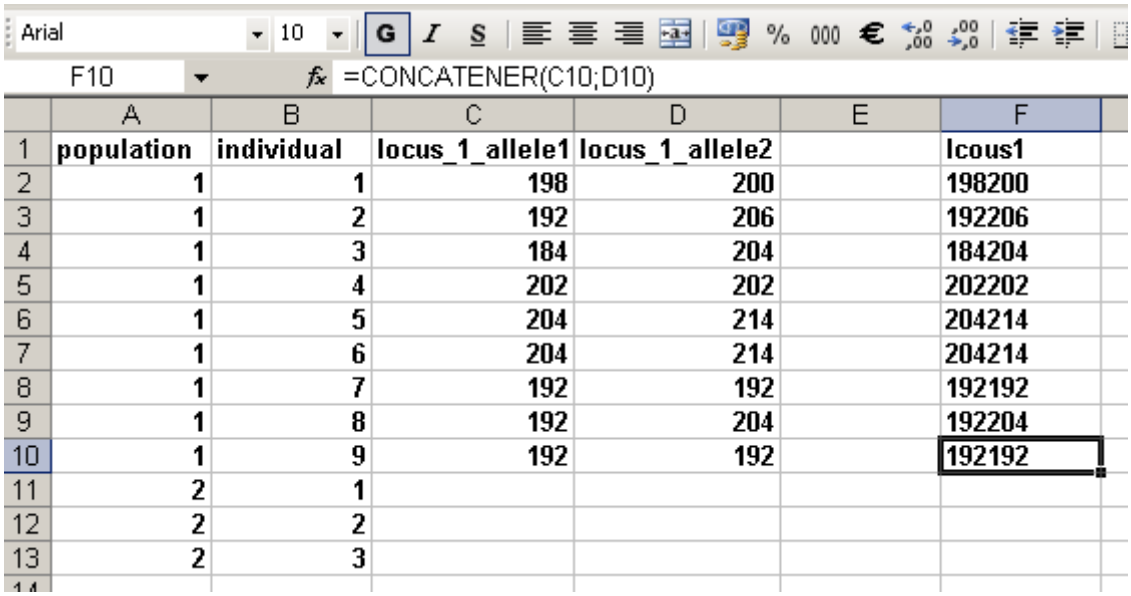
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	pop	id	b1		b2		b3		b4		b5		b6		b7		b8		b9		b10	
2	1	1	1	1	1	1	2	2	1	1	1	1	2	2	2	2	2	2	1	1	1	1
3	1	2	1	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	1	1	1	1
4	1	3	2	2	1	1	2	2	1	1	2	2	2	2	1	1	1	1	2	2	2	2
5	1	4	1	1	2	2	2	2	1	1	1	1	1	1	1	1	1	1	2	2	1	1
6	1	5	2	2	1	1	2	2	2	2	1	1	1	1	2	2	1	1	2	2	1	1
7																						

16.1.6. Notes of formatting diploid data with spread sheets

Many programmes for analysing diploid data have the bad habit (among many) of using fixed length characters for each marker. For example, our first individual with genotype 198 / 200, may need to be coded as "198200" in a single string of characters. Moreover, the same genotype in categorical coding 3 / 4 may need to be coded "0304" in a so-called **two-allele** coding, or "003004" in a **three-allele** coding. This is particularly true for the programmes Fstat and GenePop on the web. By the way, other programs may need coding as 198.200, or 198, 200, *etc*. but in general, they are handled automatically by some software (see below).

Spread sheet programmes as OpenOffice Calc or Excel handle text conversions with the CONCAT string function that can be seen in the example below.

Table 16.1–5. Example of our size type coding where two columns (one for each possible allele) have been collapsed and "concatenated" in a single text. This one is from a French version of the software and the name of the function changes a bit from language to language. For OpenOffice in English, the function is: =CONCATENATE(A1;B1), and they are accessible from the f_x button, **string** functions.



	A	B	C	D	E	F
1	population	individual	locus_1_allele1	locus_1_allele2		lcous1
2	1	1	198	200		198200
3	1	2	192	206		192206
4	1	3	184	204		184204
5	1	4	202	202		202202
6	1	5	204	214		204214
7	1	6	204	214		204214
8	1	7	192	192		192192
9	1	8	192	204		192204
10	1	9	192	192		192192
11	2	1				
12	2	2				
13	2	3				

16.1.7. Transforming data types using software

As already noted, there is not a universal data type, but some conversions can be done with available software, at least for some applications. For many programmes, there is no way around and data files must be coded manually. A small utility that we will use is the software CONVERT (Glaubitz 2003). This software can translate from a rather simple data file, to several other programmes, as shown in Figures 13.4 and 13.5:

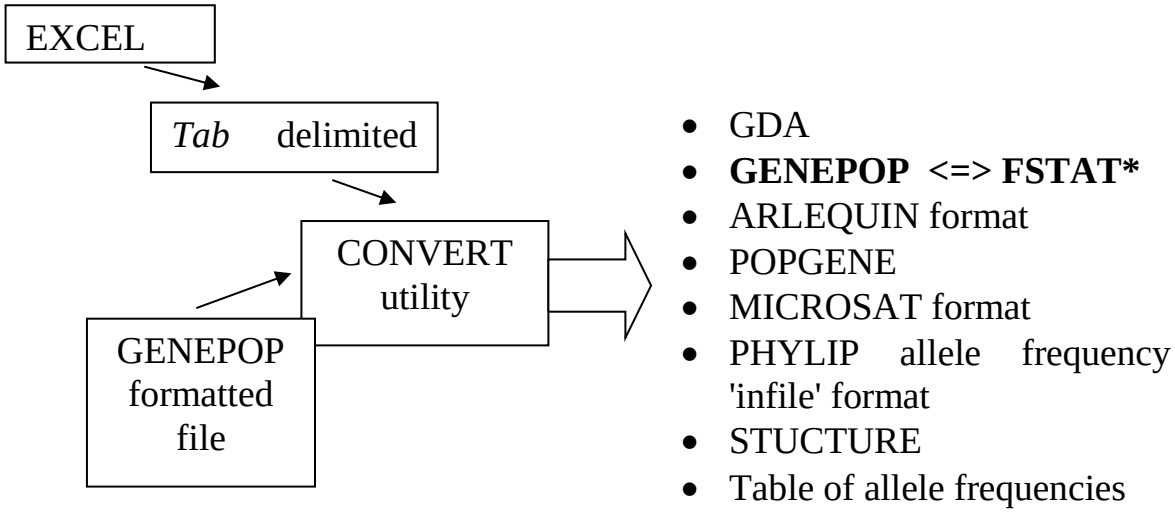


Figure 16-3. Flow chart showing the different data translation paths possible with the CONVERT utility software. Not all possibilities are here, but at least these programmes are glued together. Note, however, that these programs are almost exclusive for diploid codominant data, but some tricks can be done as explained in section 13.1.6. **FSTAT** is marked with an asterisk as is the one that we are going to use for most of the analyses, as explained in the next section.

16.1.8. The FSTAT data file

As we will use this programme mostly throughout the exercises let us explain briefly the data structure need.

For running FSTAT, it is first necessary to create an input file named FILENAME.DAT (where FILENAME is anything between 1 and 256 characters) containing the genotypic data, coded numerically, either with a 1, a 2 or a 3-digit number per allele. The file must have the following format:

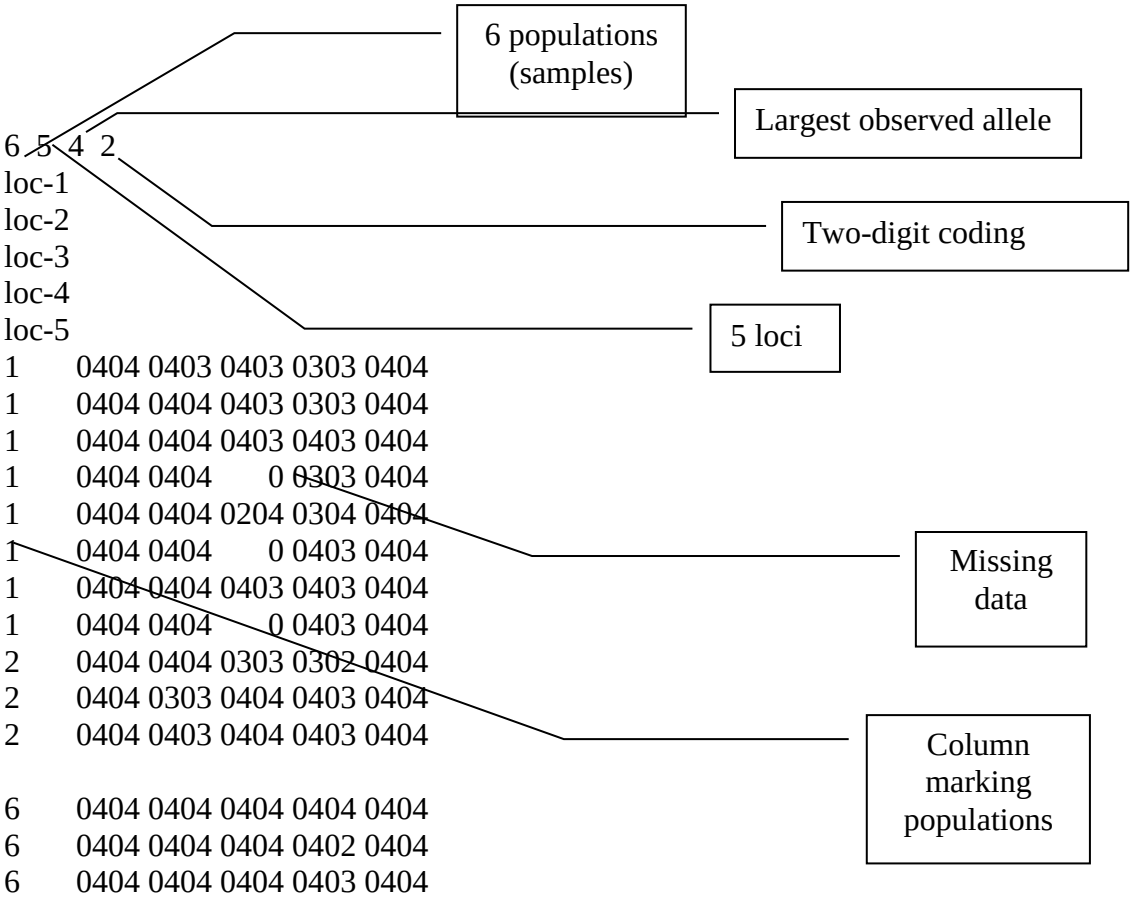
- The first line contains 4 numbers:
  1. the number of populations (here called samples) <=200
  2. the number of loci <=100
  3. the highest number used to label an allele <=999, and a
  4. data coding type: 1 if the code for alleles is a one digit number (1-9), a 2 if code for alleles is a 2 digit number (01-99) or a 3 if code for alleles is a 3 digit number (001-999).

These 4 numbers need to be separated by any number of spaces.

- Next, the name of the loci are written, one per line, and finally, the main data with first a number for each population followed by the different genotypes, each row for each individual.

- Missing data is encoded as zeros.

A data file for six populations, five loci, 4 alleles maximum and 2-digit allele coding would look then as:




```
6 5 4 2
loc-1
loc-2
loc-3
loc-4
loc-5
1      0404 0403 0403 0303 0404
1      0404 0404 0403 0303 0404
1      0404 0404 0403 0403 0404
1      0404 0404      0 0303 0404
1      0404 0404 0204 0304 0404
1      0404 0404      0 0403 0404
1      0404 0404 0403 0403 0404
1      0404 0404      0 0403 0404
2      0404 0404 0303 0302 0404
2      0404 0303 0404 0403 0404
2      0404 0403 0404 0403 0404

6      0404 0404 0404 0404 0404
6      0404 0404 0404 0402 0404
6      0404 0404 0404 0403 0404
```

16.2. Genetic diversity

Gene or genetic diversity is perhaps the central notion and motivation for conducting research in natural resources and crop improvement. If there were no biodiversity, we wouldn't have a job, and more importantly, we would probably not exist.

Evolution, or the change of heritable characters across generations (in the case of genes, it is simply the change of allele frequencies and genotype frequencies in time) can only occur if there is enough genetic variability upon which, natural and artificial selection can act. Hence, measuring genetic diversity is paramount in population genetics, and we will see that we use several complimentary approaches. First, we will see the descriptive statistics.

Allelic Richness: The first measure of genetic diversity is the number of alleles at a locus (see glossary for definitions), usually denoted *A*. The more allelic variants are found in a population, the more variable it is.

Rare Alleles: Often, we would like to mark a difference between the number of common and rare alleles. One way is to define a threshold of considering all alleles with frequencies below 0.05 as rare. These rare alleles are then considered important, and if they are unique or private to the population, we would stress them in our results. It is somewhat less used today.

Effective Alleles: Another way of estimating the number of alleles that contribute more to the diversity is by means of the effective number of alleles, denoted A_e. This measure uses the frequency of alleles to estimate the number of alleles *if* they were at the same frequency or at the maximum possible diversity, using the formula: $A_e = \frac{1}{2!\sum_i^l p_i^2}$, where p_i represents the frequency of each allele. This number can be seen also as how many numbers of individuals need to be sample before we repeat an allele. For example, typical results for microsatellite data include $A = 10$, and $A_e = 3.8$ (for example) meaning that we observed 10 alleles, but that 4 are common, and six are rare. Note that here rare is not exactly as in the previous definition, but simply that contribute less to the general diversity.

Polymorphic Bands: For dominant marker data, a straight forward measure of diversity is the percentage of polymorphic bands, which is simply the proportion of bands that present presence/absence variability. Usually they are counted with the 0.05 criterion.

Observed Heterozygosity: For diploid individuals (and polyploidy in general) this is a key measure obtained when using Co-dominant data. It is simply the proportion of individuals per population that have different recognizable alleles at a given locus and it is denote as H_o or h_o, being the former more

used for an average of many populations and the latter for a single population measure.

Expected Heterozygosity: This is the actual measure of genetic or gene diversity. It represents the probability that two alleles in a locus are different, and is usually denoted *H*, H_e or h_e. It is also known as Nei's genetic diversity as most of the gene diversity theory has been proposed by M. Nei in the 1970's. In general, it is computed as follows, although there are some variations to account for sample size or levels of inbreeding:

$$H_e = 1 - \sum_i^l p_i^2$$

where p_i represents again the frequency of each allele. The p^2 term represents the probability of sampling twice the same allele, or probability of homozygosity. Then, one minus this probability computed for all present alleles, gives us the probability of sampling two different alleles at a locus. It will be seen next, that this measure is calculated with respect to an ideal, or reference population that may, or may not have similar values as the observed heterozygosity. These deviations are considered next.

Shannon Index Diversity: The equivalent to the gene diversity, but this time cast in information theory, is the Shannon index borrowed from community ecology. Bands can be counted as we count species in lake and a global value can be calculated for a population as:

$$H = -\sum_{i}^{l} p_i \ln p_i$$

Sometimes we see this index estimated for co-dominant data. One drawback of this measure is that is not bounded, so values vary from population to population and comparisons are difficult, not as for H_e whose values are between 0 and 1.

Inbreeding: Inbreeding is both the process of reproduction between related individuals, and the result of this type of reproduction. The coefficient of inbreeding, denoted F_{is} or f_{is} or simply *f*, is a measure of consanguinity, and estimates the probability that within a locus from a given individuals, both alleles are the same, and more importantly, have originated from the same ancestor. It is measured as:

$F_{is} = (H_e - H_o)/H_e = 1 - H_o/H_e$

As evident from the above formula, the inbreeding coefficient measures a departure of genotype frequencies from a reference population (a so called Hardy-Weinberg population). When both are the same, or $H_o = H_e$, the inbreeding coefficient is 0, and we would say that no significant departures from HW were observed.

Significant deviations from HW, i.e., f_{is} significantly greater than zero, can arise for a number of reasons that are not mutually exclusive, mainly:

- *Small population* size that entails the loss of heterozygotes just by chance (genetic drift) and increases the probability of mating with related individuals;
- *Non-random mating* that favours the replication of the same genotypes in the population;
- *Selfing* (plants and certain snails), which is a form of non-random mating
- *Lack of external gene flow*, without migration, alleles will be fixed just by chance in small, isolated populations.

Testing for significant inbreeding is performed with different tests (i.e., fisher's exact tests), but many programmes rely in permutation tests to find a numerical solution for it. For example, FSTAT reshuffles alleles within loci to create a null distribution of possible f_{is} values from the data, and then compares if the observed value is at one or the other extreme of this distribution that is centred approximately at zero. If the observed f_{is} is in one of the extremes that contain 2.5 % of the simulated data (a 5% two-sided test), we would conclude that the f_{is} is true value greater than zero, and not a random result.

16.3. Genetic structure

In section 11.2, we saw a series of descriptive genetic diversity parameters, that summarizing are: A, A_e, H_o, H_e, and f_{is}. When we have two or three populations, comparisons are feasible, but things can be more complicated for more samples. Moreover, we could begin to loose information, even with few populations, because the measures of inbreeding, for example, are performed with population-specific data that does not tell us anything about the relative value of diversity, or inbreeding of *all* populations.

As a definition, genetic structure refers to the *non*-random distribution of genetic diversity in space and time.

16.3.1. Nei's population genetics parameters: G_{st} family

Casting our question in terms of H's or genetic diversities only, we might ask how is the total genetic diversity related to the average sup-population diversity? In other words, has the total population more information than that existing in a single population? Or, are all populations the same?

To answering these questions, Nei developed in 1972 a synthetic parameter called G_{st}. This parameter takes the value of zero, if all sub-populations contain the same information as the total population, and greater than zero and up to one (rarely achieved), if any of the sub-populations contains levels of diversity that are not distributed at random among the sup-populations.

Its computation is rather straight forward and follows the equation:

$$G_{st} = (H_t - H_s)/H_t = 1 - H_s/H_t$$

Where H_t is the total population diversity (computed from the average allele frequencies from all subpopulations) and H_s is the average within population diversity computed for each single population. It is clear that if both values are the same, G_{st} approaches zero. If not, if H_t is much larger than H_s, we would say that the distribution of genetic diversity is not random, or is structured.

16.3.2. Sewall Wright's F-statistics

If instead of thinking of diversity, but inbreeding, or better correlation of alleles within ***I***ndividuals, ***S***ubpopulations and the ***T***otal population, a set of relationships can be deduced for the different levels at which genes occur (individuals, subpopulations and the total population, of course). Thus, the inbreeding coefficient that we saw earlier for a single population can be "scaled" to different levels of population organization and different inbreeding coefficient can be used. Thus, we can ask ourselves about of:

- the correlations of gametes within individuals relative to the subpopulation, or $F_{IS;}$
- the correlations of gametes within individuals relative to the total population, or $F_{IT;}$
- the correlations of gametes within subpopulations relative to the total population, or $F_{ST.}$

If any of these correlations is >> 0, it means that the probability of finding two identical alleles is stronger in the subunit (individual or subpopulation) than in the reference population (subpopulations and total population). Note that in principle, all this values are between zero and one, closeness to one meaning fixation of alleles at the particular scale. Note also that capital letters have been used to distinguish these parameters from single-population parameters. They are related by the expression:

$$(1 - F_{IT}) = (1 - F_{IS}) \ldots (1 - F_{SR}) (1 - F_{...}) \ldots (1 - F_{IS})$$

Where F_{SR} and $F_{..}$ have been introduced between F_{IS} and F_{ST} to denote that population structure can be more complex and include regions, watersheds, *etc*.

The two most common used statistics are F_{IS} and $F_{ST,\ but}$ F_{IT} has been overshadowed by the rest. Note also that for Nei's *G*-statistics, there are equivalent G_{is}, G_{it}, but are less and less used.
F_{st} is commonly regarded as *the* population structure parameter that if significantly greater than zero indicates that diversity (or inbreeding) is not randomly distributed. Several other parameters, however, have been proposed by different authors and the list grows almost every year. We will highlight some of the most used:

- Weir's and Cokerham's Θ (theta), also now as the co-ancestry coefficient. Reputedly more robust to sampling variation than the basic F_{ST}.
- Excoffier's et al. Φ (*Phi*)-statistics, that are analogous to F_{ST}, but based on variance components analyses.
- R_{ST} (with its estimator ρ (*rho*)) that uses the actual microsatellite size to estimate the genetic structure parameter. Note, if microsatellites are coded as allelic states, we would be estimating *Phi*-statistics.
- N_{ST}, analogous to the others, but for sequencing data (seldom used, more of a theoretical value).

16.4. Population and individual divergence and phylogenetic trees

So far, we have seen that a complete description of genetic diversity entails first, the estimation of various descriptive parameters for each subpopulation, and then, the use of synthetic values that will allow us to tell if genetic diversity is distributed at random or not (*i.e.*, Fst >> 0). However, can we tell apart which population(s) is actually producing this structure? Which populations are more divergent than others, and in which direction?

These questions are then answered by using a divergence analysis based on genetic distances. Strictly speaking, unless we use particular methods that can validate a direction of

evolutionary changes (uses of out-groups, identification of ancestral characters or states, *etc*.) we would be doing phenetic analysis. This means that we are able to pinpoint out the separation of populations, or individuals, but we cannot know which end of the "phylogenetic" tree precedes the rest. In crop improvement, however, this is not usually a big problem as groups are arbitrarily chosen and what matters is what is different from the others.

Similarly as for genetic structure (see section 3), there exist several ways of estimating individual or population genetic distances, but the procedure is always the same:

- Define a distance metric.
- Calculate distances among groups or among individuals (results are usually stored in a pairwise matrix of genetic distances whose diagonal is zero). If possible, bootstrap loci or individuals (*i.e*., resample information to validate observed results) to get a support for the branches of the tree.
- Visualize the resulting distance using a particular algorithm.

In our case, the two most used algorithm for visualizing distances among groups are UPGMA (Un-weighted Pair Group Method with Arithmetic Mean) and Neighbor-joining. The former is the simplest method of tree construction. It was originally developed for constructing taxonomic phenograms, *i.e*. trees that reflect the phenotypic similarities between species, but it can also be used to construct phylogenetic trees if the rates of evolution are approximately constant among the different lineages. The latter, Neighbor-joining (Saitou and Nei, 1987) is a method that is related to the cluster method but does not require data whose lineages have diverged by equal amounts.

Common genetic distances include:

- Nei's genetic distance (Nei, 1972);
- Cavalli-Sforza chord measure (Cavalli-Sforza and Edwards, 1967)l
- Reynolds, Weir, and Cockerham's genetic distance (1983).

These types of analyses are well handled by the set of program PHYLIP, and also by POPULATIONS, although any software that can produce a distance matrix will be useful for producing a tree. Testing of the branches and tree structure, however, is a delicate task and is mostly the domain of phylogenetics instead of population genetics, although the two fields overlap.

16.5. Web resources and software – non-exhaustive

FSTAT:
http://www2.unil.ch/popgen/softwares/fstat.htm

Pros: General purpose diploid analysis software with not so difficult data file. Nice interface, very good help files and handles most of the necessary analyses. Output files are also good, almost ready to use.

Cons: doesn't perform nested Fst analyses. Does not report per population Ho(!).
GenePop on the Web:
http://wbiomed.curtin.edu.au/genepop/

Pros: Frequently updated, includes many tests for the significance of inbreeding, available everywhere through the web.

Cons: doesn't perform nested Fst analyses either. Output tables are awful and confusing. Ho is reported not as a fraction, but as the count (observed and expected) of heterozygote individuals.

Arlequin:

Pros: so far, the most comprehensive software devoted for population genetics. Does handle nested Fst (or hierarchical AMOVA's). Excellent manual that serves as a summary of population genetic methods, highly recommended!

Cons: one of the worst data file format ever! This has been circumvented by the automatic translation by other software, to certain limits. Interface apparently simple, but results are mixed with original data files, becoming confusing after many runs.

AFLPsurv:
http://www.ulb.ac.be/sciences/lagev/aflp-surv.html

Pros: I have yet to see a dominant marker program that convinces me, but this is a workable one. Includes many genetic distances and calculates genetic diversity.

Cons: Bootstrapping for individuals is restricted as it is population oriented software.

PHYLIP:
http://evolution.genetics.washington.edu/phylip.html

Pros: this is a collection of programs, and is somewhat the dean of phylogenetic analyses. Has been overshadowed by PAUP, but as free software is a good starting point, and although methods are somewhat outdated, the implementation is serious.
Cons: as said, somewhat outdated, but good for most applications.

TreeView:
http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Pros: small and effective program for visualizing trees constructed in the PHYLIP format (i.e., out files from NEIGHBOUR, for example).

Cons: large trees appear sometimes not so well, no possibility of editing trees.

Populations:

Pros: very good collection of genetic distances for codominant markers. It can deal with dominant marker data if we use the 22-11 coding. Produces tree files directly observable with TreeView and accepts GenePop data files.

Cons: often times it crashes unexpectedly possibly because of missing data or repeated individual names within populations.

RstCalc:
http://helios.bto.ed.ac.uk/evolgen/rst/rst.html

Pros: good programme for estimating Rst.

Cons: Data file is not difficult, but could be simpler. It does not handle nested Rst.

CONVERT:
http://www.agriculture.purdue.edu/fnr/html/faculty/Rhodes/Students%20and%20Staff/glaubitz/software.htm

Pros: little programme that uses a simple excel file that can be translated into other software, including GenePop and Arlequin.

Cons: does not support FSTAT, so passing through GenePop is necessary.

COOK BOOK
FORMATTING POPULATION GENETIC DATA

(1). Step 1:
Scoring the data
Record data in excel file and transform as necessary. For the programme populations, to be used in our demonstration, a 2-digit formatting is required.

- For dominant markers, ISSR, AFLP, IRAP or others scored as present or absent, *i.e.* 1 or 0, transform as follows:

  Manually select all data input (taking care not to select the names of the individuals, populations or loci)
  First, replace all '1' with '22'
  Second, replace all '0' with '11'
  (At this point it is helpful to check for missing data)
- For codominant markers e.g. SSR data are already scored as 2 digits so no need for transformation
- For mixture of dominant and codominant markers, transform the dominant to codominant by scoring as 2-digit

(2) Step 2:
Formatting the data for populations programme

- Insert a new row between the header row (*i.e.* A, B, C, …) and first row such that newly inserted row becomes row no. 1 and then do the following in the new row (*i.e.* Row No. 1):
  - First column: type in the number of populations or samples
  - Second column: type in the number of loci or markers
  - Third column: type in the highest number used to label an allele
  - Fourth column: data coding type [*1 if the code for alleles is one digit number (1-9); 2 if code for alleles is a 2 digit number (01-99) or a 3 if code for alleles is a 3 digit number (001-999)*]
- Insert another row between now rows 2 and 3 and do the following:
  - In the first column, type '**pop**'

(3). Step 3:
Formatting the data as a "tab delimited text file"

- Select all entries by highlighting (starting from cell A1X1 to the end of the data entries)
- File > Save as > text (tab delimited) (*.txt) > OK > Yes.
- Save on same disk and folder as the Populations.exe file (To run the programme, the text file (.txt) must be in same folder as 'Populations.exe'.

(4). Step 4:

- Formatting in NOTEPAD:
- Open NOTEPAD
- From File menu, locate the saved .txt file, open file

- Put cursor in front of first locus and hit backspace so that it is now in the first column, second row
- Highlight all entries by select all in Edit menu
- Cut (the entries)
- Paste in Word
- Select All
- Edit > Replace > In "*find what*" box type "**^t**" and in "*replace with"* box, hit the space bar once. Select 'replace all' option. All the tabs have been replaced. (It helps to have the paragraph icon on in order to see that there are only single spaces).
- Delete the dots (after the figures in the first row and after 'pop' and insert comma each sample name. Make sure that there are no spaces within a sample name.
- Select all entries
- Cut (the entries)
- Paste again in NOTEPAD
- Put the cursor in front of the first data in each row and hit backspace (the space between the 'comma' after the sample name and the first score is deleted)
- Save (Use a simple file name – one word).
- Save the .**txt file** in the same folder as the Programme, 'Populations.exe'.

(4). Step 4:
Running the programme
Open program and choose sequentially by entering the corresponding numbers and hitting 'Enter':

- Compute individuals distance + tree (when data has only one population) – No. 1
- Type the exact name of .txt file from last saving in the space provided. The '.txt' extension must be included in the name. The name is also case sensitive.
- Phylogenetic tree of individuals with bootstraps on locus – No. 3
- Nei's standard genetic distance, Ds (1972) – No. 2
- UPGMA – No. 1
- 10000
- Enter desired name for output file with '.tre' extension
- Wait for the programme to finish running. The output file with the '.tre' extension is now deposited in the same folder as the programme, 'Populations.exe'
- Double click on the output file with the '.tre' extension in order to see the resulting dendrogram.

16.6. References

Cavalli-Sforza, L. L.; Edwards, A. W. F., 1967: Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. 18, 233-257.

Chakraborty R and Danker-Hopfe H, 1991. Analysis of population structure: A comparative study of different estimators of Wright's fixation indices. In 'Statistical Methods in Biological and Medical Sciences.' Ed C.R. Rao and R. Chakraborty, Elsevier Science Publishers.

Cockerham CC, 1969. Variance of gene frequencies. Evolution. 23:72-84.
Cockerham CC, 1973. Analysis of gene frequencies. Genetics. 74:679-700.
Cockerham CC and Weir BS, 1993. Estimation of gene-flow from F-statistics. Evolution. 47:855-863.
El Mousadik A and Petit RJ, 1996. High level of genetic differentiation for allelic richness among populations of the argan tree [Argania spinosa (L.) Skeels] endemic to Morocco. Theor. Appl. Genet. 92:832-839.
Excoffier L 2001. Analysis of population subdivision. In Handbook of statistical genetics, Balding, Bishop & Cannings (Eds) Wiley & Sons, Ltd.
Fisher R, 1954. Statistical Methods for Research Workers. 12th Edition, Oliver & Boyd, Edinburgh. 356pp.
Goodman SJ, 1997. Rst Calc: a collection of computer programs for calculating estimates of genetic differentition from microsatellite data and a determining their significance. Molecular Ecology 6: 881-885.
Glaubitz, J.C. (submitted) CONVERT: A user-friendly program to reformat diploid genotypic data. Molecular Ecology.
For commonly used population genetic software packages. Molecular Ecology Notes.
Goudet J, 1995. FSTAT (vers. 1.2): a computer program to calculate F-statistics. J. Hered. 86: 485-486.
Goudet J, Raymond M, Demeeus T and Rousset F, 1996. Testing differentiation in diploid populations. Genetics. 144:1933-1940.
Hartl DL, Clarck AG (1997) Principles of Population Genetics. Third Edition. Sinauer Associates.
Nei, M. (1972) Genetic distance between populations. Am. Nat. 106:283-292.
Nei M, 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA. 70:3321-3323.
Nei M, 1988. Molecular Evolutionary Genetics. Columbia University Press, New York.
Nei M and Chesser RK, 1983. Estimation of fixation indices and gene diversities. Ann. Hum. Genet. 47:253-259.
Pamilo P, 1984. Genotypic correlation and regression in social groups: multiple alleles, multiple loci and subdivided populations. Genetics. 107:307-320.
Petit RJ, El Mousadik, A and Pons O, 1998. Identifying populations for conservation on the basis of genetic markers. Conservation Biology. 12:844-855.
Queller DC and Goodnight KF, 1989. Estimating relatedness using genetic markers. Evolution. 43:258-275.
Raymond M. & Rousset F, 1995. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. Journal of Heredity, 86, 248-249.
Raymond M and Rousset F, 1995. An exact test for population differentiation. Evolution. 49:1280-1283.
Reynolds J, Weir BS and Cockerham CC, 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. Genetics. 105:767-779.
Rousset F, 1996. Equilibrium Values of Measures of Population Subdivision For Stepwise Mutation Processes. Genetics 142:1357-1362.
Rousset F, 1997. Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145:1219-1228.

Saitou, N and M Nei, 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.
Slatkin M, 1993. Isolation by distance in equilibrium and non-equilibrium populations. Evolution 47:264-279.
Slatkin M, 1995. A measure of population subdivision based on microsatellite allele frequency. Genetics. 139:457-462.
Slatkin M and Barton NH, 1989. A comparison of three methods for estimating average levels of gene flow. Evolution 43:1349-1368.
Sokal RR and Rohlf FJ, 1981. Biometry. 2nd Edition. Freeman & Co.
Weir BS and Cockerham CC, 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.
Weir BS, 1996. Genetic data analysis II. Sinauer Publ., Sunderland, MA.
Whitlock MC and McCauley D, 1999. Indirect measures of gene flow and migration: Fst<>1/(4Nm+1). Heredity. 82: 117-125.
Wright S, 1969. Evolution and the genetics of populations. Vol. 2. The theory of gene frequencies. University of Chicago Press.

16.7. Some key concepts

Alleles: All possible forms of a gene.
Gene: A unit of inheritance, a non-recombining segment of DNA. A given location on a chromosome
Genotype: The combination of the two homologous alleles carried on the two chromosomes of a diploid individual at a given locus.
Haplotype: A particular combination of alleles at different loci on a chromosome.
Heterozygosity: The probability of an individual to have two different alleles at a given locus (the probability of being heterozygote).
Homozygosity: The probability of an individual to be homozygote at a given locus.
Homozygote: The fact that an individual has two identical alleles at a given locus.
Locus: A given location on a chromosome, a non-recombining segment of a chromosome (usually interchanged with gene)
Phenotype: The visible (physical) state of an individual. The relation between the genotype and the phenotype can be complex and will usually depend on the degree of dominance and the interaction of different alleles at a single or multiple loci.
Polymorphism: the fact that there exist different alleles at a given locus in a population.
Population: A group of interbreeding individuals living together in time and space. It is usually a subdivision of a species.
Sample: A collection of individuals or of genes drawn from a population.

16.8. Equations

$$p = P + H/2 \qquad \widehat{H}_E = \frac{2n}{2n-1}\left(1 - \sum_{i=1}^{k} p_i^2\right) \qquad H_e = 2pq - \frac{2Spq}{2-S}$$

$$\frac{Q_f}{Q} = \frac{q^2 + fpq}{q^2} = 1 + \frac{fp}{q} \qquad H = 2pq(1-f) \qquad q_2 = \frac{1}{2}(q_f + q_m)$$

$$var(p') = \frac{pq}{2N} \qquad H_t = 2p_0q_0 + A^t(Q_0 - 2p_0q_0)$$

$$P(X = r) = \frac{2N!}{r!\,(2N-r)!}p^r(1-p)^{2N-r} \qquad f = \sum_{i=1}^{m}\left(\frac{1}{2}\right)^{N_1}\left(1 + f_{CA(i)}\right)$$

$$f(t) = 1 - \left(1 - \frac{1}{2N}\right)^t\left(1 - f(0)\right) \qquad H(t) = \left(1 - \frac{1}{2N}\right)^t H(0)$$

$$t \approx -\ln\left(\frac{H(t)}{H(0)}\right)2N \qquad var\big(p(t)\big) = p_0q_0\left(1 - \left(1 - \frac{1}{2N}\right)^t\right)$$

$$D = \frac{t}{2N} \approx -\ln\left(1 - \frac{\mathrm{var}(p)}{\bar{p}\bar{q}}\right) \qquad N_e = \frac{4N_mN_f}{N_{f+}N_m} \qquad N_e\frac{N\bar{k} - 1}{\bar{k} - 1 + \frac{\mathrm{var}(k)}{\bar{k}}}$$

$$\frac{1}{N_e} = \frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{N(i)} \qquad N_e = \frac{N}{1+f} \qquad N_e = 4\pi\sigma^2 d$$

$$\overline{H_S} = 2\bar{p}(1-\bar{p}) - 2\mathrm{var}(p) \qquad F_{ST}\frac{var(p)}{\bar{p}(1-\bar{p})} \qquad (1-x)^y \approx e^{-xy}\ \ if\ x < 0$$

$$D = \frac{t}{2N} \approx -\ln(1 - F_{ST}) \qquad (1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST}) \qquad \bar{F}_{IT} = \frac{\bar{H}_T - \bar{H}_0}{\bar{H}_T}$$

$$\bar{F}_{ST} = \frac{\bar{H}_T - \bar{H}_S}{\bar{H}_T} \qquad \bar{F}_{IS} = \frac{\bar{H}_S - \bar{H}_0}{\bar{H}_S} \qquad F_{ST} \approx \frac{1}{4Nm + 1}$$

$$F_{ST} \approx \frac{1}{4Nm\left(\frac{k}{k-1}\right)^2 + 1} \qquad N_e \approx \frac{kN}{1 - F_{ST}} \qquad p' = p\frac{w_1}{\bar{w}}$$

$$q' = q\frac{\overline{w_2}}{\bar{w}} \qquad \bar{w} = p^2 w_{11} + 2pqw_{12} + q^2 w_{22} \qquad \overline{w_2} = qw_{22} + pw_{12}$$

$$\Delta q = q\frac{\overline{w_2} - \bar{w}}{\bar{w}} \qquad q_e = \frac{u}{u+v} \qquad p_t = (1-u)^t p_0$$

$$u(p) = p_0 = \frac{1}{2N} \qquad u(q) = 1 - u(p) = 1 - \frac{1}{2N} \qquad T_1(p) = 4N_e$$

$$T_0(p) = 2\frac{N_e}{N}\ln(2N) \qquad F_e \approx \frac{1}{4Nu+1} = \frac{1}{\theta+1} \qquad H_e = \frac{\theta}{\theta+1}$$

$$u\left(\frac{1}{2N}\right) \approx 2s \qquad T_1(p) = \frac{2}{s}\ln(2N) \qquad r = \frac{1}{u}$$

$$r = \frac{1}{4Nus} \qquad E(k|n,\theta) = 1 + \theta\sum_{i=1}^{n-1}\frac{1}{\theta+i} \qquad K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}d\right)$$

17. APPENDICES

17.1. General DNA extraction techniques

17.1.1. Phenol/chloroform extraction

NOTE: Wear gloves, goggles, and lab coat at all times for safety and to prevent contamination of your preparations.

Removes protein from DNA preparations. Advisable for example if A_{260nm}: A_{280nm} (from the spectrophotometer readings) of the DNA are below 1.6. Phenol extraction requires subsequent ethanol precipitation of the DNA.

Phenol: freshly distilled and equilibrated with 20 % 0.5 M Tris-Base. Prepare a mixture of phenol/chloroform/isoamylalcohol (PCI) (25:24:1).

NOTE: Use caution as phenol is toxic.

1. The DNA sample is mixed with an equal volume of PCI, vortexed, and centrifuged for about 5 minutes. Remove the upper aqueous phase avoiding contamination with protein from interphase and transfer it to a fresh reaction tube.
2. Remaining traces of phenol in the aqueous phase are extracted with 1 volume of chloroform/isoamylalcohol (24:1). Vortex and centrifuge for 5 minutes. Transfer the upper phase carefully to a fresh reaction tube.

17.1.2. Ethanol precipitation

NOTE: Wear glasses at all time for safety.

1. Determine volume of the sample, add 0.1 volume 3 M sodium acetate and 2.5 volumes cold ethanol (96%). Mix well and leave at -20°C for 2 hours.
2. Centrifuge for 15 minutes (in microcentrifuge at >12,000 rpm), preferably at 4°C.
3. Carefully remove ethanol and wash pellet with cold 70% ethanol to remove salt from the sample – centrifuge for 5 minutes.
4. Dry DNA pellet in vacuum centrifuge or air dry in flow bench.
5. Dissolve DNA in TE buffer or sterile double distilled H_2O (ddH_2O).

17.1.3. Solutions

- 1.5 x CTAB extraction buffer (1 liter):

CTAB	15.0 g	
1 M Tris (pH 8.0)	75 ml	
0.5 M EDTA	30 ml	
NaCl		61.425 g

ddH_2O	to 1 litre

- 10% CTAB (1 litre)

CTAB	100 g
NaCl (0.7M)	40.95 g
ddH_2O	to 1 litre

- β-mercaptoethanol,
- Chloroform:isoamylalcohol (24:1),
- Isopropanol,
- Ethanol 96% and 70%
- sodium acetate (3 M)
- TE buffer

10 mM Tris HCl
1 mM EDTA (pH 8.0)

17.2. Polymerase chain reaction protocol

The polymerase chain reaction (PCR) is basically a technique for *in vitro* amplification of specific DNA sequences by the simultaneous primer extension of complementary DNA strands. The principle of primer extension is illustrated in Figure A.2.1 for one DNA strand. The primer binds to its complementary sequence of the single stranded target DNA and the polymerase extends the primer in 5' - 3' direction by using the complementary DNA as a template. For a PCR reaction, two primers are used, one binding to the "lower" strand (forward primer) and one binding to the "upper" strand (reverse primer). Thus, the requirements for the reaction are: template DNA, oligonucleotide primers, DNA polymerase, deoxynucleotides (to provide both energy and nucleosides for DNA synthesis), and a buffer containing magnesium ions. In general the DNA sequence of both ends of the region to be amplified must be known to be able to synthesize proper primer oligonucleotides. The PCR reaction is a cyclic process, which is repeated 25 to 35 times. One cycle consists of three basic steps with characteristic reaction temperatures:

1. Denaturation of the double stranded DNA to make the template accessible for the primers and the DNA polymerase (94°C, 30 seconds).
2. Annealing of primers to complementary sequence on template (between 45 and 60°C, depending on the primer sequences, 30 seconds).
3. Extension of primers by DNA-polymerase (72°C - the optimum temperature of *Taq* DNA-polymerase -, 1 minute per kilobase of template to be amplified).

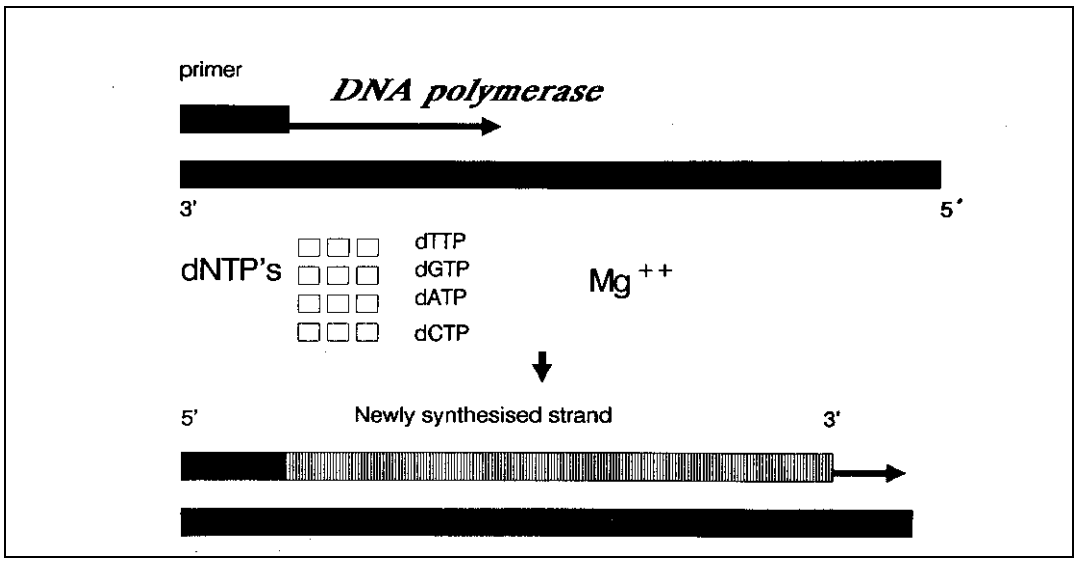


Figure A.2.1. Primer extension. DNA polymerase extends a primer by using a complementary strand as a template (McPherson *et al.*, 1991)~~(McPherson *et al.*, 1991)~~.

By multiple repetition of this cycle the number of template molecules increases. This result in exponential amplification of the DNA sequence that is bordered by the two primers used (Figure A.2.2).

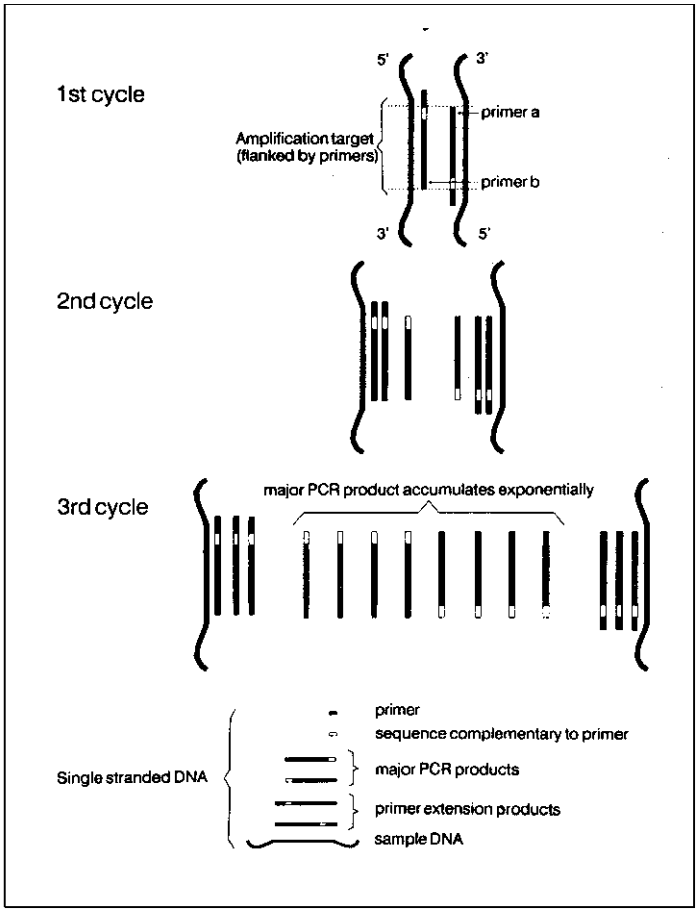


Figure A.2.2. Schematic diagram of PCR. By using primer pairs 'a' and 'b' (short black lines) annealed to complementary strands of DNA (long black lines), two new strands (shaded lines) are synthesized by primer extension. If the process is repeated, both the sample DNA and the newly synthesized strands can serve as templates, leading to an exponential increase of product which has its ends defined by the position of the primers (McPherson *et al.*, 1991)~~(McPherson *et al.*, 1991)~~.

Successful performance of a PCR experiment is dependent on a number of different factors; some of them have to be determined empirically.

- The selection of the primers is a very important step. They should be long enough to be specific, not anneal against themselves by folding (avoid palindromic sequences), nor should the forward primer anneal with the reverse primer. Furthermore the G/C content of the primers should be similar and they should have similar melting temperatures (T*m*). Several computer programs are available on the Internet to help to find the best primer pairs for a given sequence. Try the addresses below- submit the DNA sequence and some required parameters and you will get a list of possible primers:

http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi
http://genome-www2.stanford.edu/cgi-bin/SGD/web-primer
http://www.nwfsc.noaa.gov/protocols/oligoTMcalc.html

- The annealing temperature must be determined empirically and is dependent from the T*m*'s of the primers. A rule of thumb (Wallace rule) provides a first order approximation for T*m* of oligonucleotides that have 20 bases or less:

$$Tm = 2°C\ (A + T) + 4°C\ (C + G)$$

The annealing temperature is a few degrees lower than T*m*.

- PCR is extremely sensitive! Thus contamination of samples and solutions with minimal amounts of foreign DNA, or the wrong PCR programme can result in unspecific PCR products. Always include controls without template DNA in order to check if there is any contamination in your nucleotides, primers, etc.

A typical PCR experiment is given in the table below. In the FAO/IAEA course, PCR was demonstrated by amplifying a 1050 bp sequence of the rice retrotransposon Tos 17 accession number D88394:

Forward Primer 1 (100 pmol/µl):
Reverse Primer 2 (100 pmol/µl):
Reaction volume: 50 µl

Stock solutions	**µl**	**Final conc./amount**
10 x PCR buffer (15 mM $MgCl_2$)	5.0 µl	1 x PCR buffer (1.5 mM $MgCl_2$)
Primer 1 (100 pmol/µl)	0.5 µl	1 pmol
Primer 2 (100 pmol/µl)	0.5 µl	1 pmol
dNTP mix (10 mM)	1 µl	0.2 mM
DNA template (100 ng/µl)	1 µl	100 ng
Taq DNA Polymerase (5 U/µl)	0.5 µl	2.5 U
H_2O	41.5 µl	-

NOTE: It is very important to prepare a master mix corresponding to the number of desired samples that contains all the reagents except for the template DNA. Mix well and add the appropriate amount of the master solution to single reaction vials containing the individual template DNA samples you wish analysed. This procedure significantly reduces the number of pipetting steps, avoids errors derived from pipetting small amounts of liquid, and finally ensures that every tube contains the same concentrations of reagents.

For amplification of the Tos17 sequence the PCR machine was programmed as follows:

Step 1	Initial denaturation	94°C	(4:00 minutes)
Step 2	Denaturation	94°C	(0:30 minute)
Step 3	Primer annealing	56°C	(0:30 minute)
Step 4	Primer extension	72°C	(1:10 minutes)
Step 5	Cycling	Repeat steps 2-4	29 times
Step 6	Final extension	72°C	(6:00 minutes)
Step 7	Hold	4°C	(hold)

NOTE: The PCR programme can vary from primer to primer set and species to species with the annealing temperature being the most variable step.

17.2.1. References

McPherson, M., P. Quirke, and G Taylor, 1991. PCR: A Practical Approach. Oxford University Press, New York.

17.3. Plant genome database contact information

Table 17.3–1 Taken from an IAEA-TECDOC on "Radioactively Labelled DNA Probes For Crop Improvement" VIENNA SEPTEMBER 6-8, 1999).

DATABASE	CROPS	CURATOR	E-MAIL ADDRESS	DATABASE ADDRESS
AAtDB	*Arabidopsis*	David Flanders	flanders@genome.stanford.edu	http://genome-www.stanford,edu/Arabidopsis/
Alfagenes	Alfalfa (*Medicago sativa*)	Daniel Z. Skinner	Dzolek@ksu.ksu.edu	http://naaic.org/
Bean Genes	*Phaseolus* and *Vigna*	Phil McClean	mcclean@beangenes.cws.ndsu.nodak.edu	http://probe.nalusda.gov:8300/cgi-bin/browse/beangenes
ChlamyDB	*Chlamydomonas reinhardtii*	Elizabeth H. Harris	chlamy@acpub.duke.edu	http://probe.nalusda.gov:8300/cgi-bin/browse/chlamydb
CoolGenes	Cool season food legumes	Fred Muehlbauer	muehlbau@wsu.edu	http://probe.nalusda.gov:8300/cgi-bin/browse/coolgenes
CottonDB	*Gossypium* species	Sridhar Madhavan	msridhar@tamu.edu	http://probe.nalusda.gov:8300/cgi-bin/browse/cottondb
GrainGenes	Wheat, barley, rye and relatives	Olin Anderson	oandersn@pw.usda.gov	http://probe.nalusda.gov:8300/cgi-bin/browse/graingenes
MaizeDB	Maize	Mary Polacco	maryp@teosinte.agron.missouri.edu	http://www.agron.missouri.edu/
MilletGenes	Pearl millet	Matthew Couchman	Matthew.Couchman@bbsrc.ac.uk	http://jiio5.jic.bbsrc.ac.uk:8000/cgi-bin/ace/search/millet.
PathoGenes	Fungal pathogens of small-grain cereals	Henriette Giese	h.giese@risoe.dk	http://probe.nalusda.gov:8300/cgi-bin/browse/pathogenes
RiceGenes	Rice	Susan McCouch	srm4@cornell.edu	http://genome.cornell.edu/rice/
RiceGenome Project	Rice			http://www.staff.or.jp
SolGenes	*Solanaceae*	Molly Kyle	mmk9@cornell.edu	http://genome.cornell.edu/solgenes/welcome.html
SorghumDB	*Sorghum bicolor*	Russel Kohel/Bob Klein	nus6389@tam2000.tamu.edu	http://probe.nalusda.gov:8300/cgi-bin/browse/sorghumdb
Soybase	Soybeans	David Grant	dgrant@iastate.edu	http://129.186.26.94/
TreeGenes	Forest trees	Kim Marshall	kam@s27w007.pswfs.gov	http://dendrome.ucdavis.edu/index.html
National Center for Genome Resources	Various			http://www.ncgr.org/

17.4. Acronyms of chemicals and buffers

AMPPD	4-Methoxy-4-(3-phosphatephenyl)spirol(1,2-dioxetan-3,2'-adamantan)
BCIP	5-Bromo-4-chloro-3-indolyl phosphate
CSPD®	Chemiluminescence substrate (a registered trademark of Tropix Inc., USA)
CTAB	Hexadecyltrimethylammonium bromide
ddH_2O	Double distilled water
DIG	Digoxygenin
N_2 liquid	Liquid nitrogen
NBT	Nitro blue tetrazolium
PCI	Phenol/chloroform/isoamylalcohol (25:24:1)
SDS	Sodium dodecyl sulphate
SSC	Saline-sodium citrate buffer
TBE	Tris-borate-EDTA buffer
TE	Tris-EDTA buffer
TEMED	N,N,N',N'-tetramethylenediamine
TRIS	[Tris(hydroxymethyl)aminomethane]